ABX-Discriminability Measures and Applications

Thèse de doctorat de l'Université Paris VI, spécialité Cerveau — Cognition — Comportement;

Préparée par Thomas Schatz;

Au sein du Laboratoire de Sciences Cognitives et Psycholinguistique (ENS/EHESS/CNRS) et de l'équipe-projet SIERRA (ENS/INRIA/CNRS);

Sous la direction d'Emmanuel Dupoux et de Francis Bach;

Défendue le 29 Septembre 2016, devant un jury présidé par Christian Lorenzi et composé de:

Daniel Swingley, Jean-Luc Schwartz, Ludovic Denoyer, Martine Adda-Decker, Francis Bach et Emmanuel Dupoux.

ABX-Discriminability Measures and Applications

Thomas Schatz

July 11, 2016

Contents

Acknowledgements

I would like to thank my thesis directors Emmanuel Dupoux and Francis Bach for their disponibility and patience throughout the completion of this thesis, for their invaluable advice, as well as for the example of scientific exigence and intellectual honesty they provided. I would like to thank Emmanuel Dupoux in particular for the time he invested and for the many inspiring discussions we had over the course of these five years. I would also like to thank all the people I scientifically interacted with, at the LSCP and beyond, and in particular the members of the BOOTPHON team for the many interesting exchanges and cheerful moments we had over the years. Finally, I would like to thank my friends, family and girlfriend for their unwavering patience and support.

Introduction

The starting point for this thesis was the problem of modeling phonetic category acquisition in infancy. Roughly speaking phonetic category acquisition refers to the process by which infants during their first year of life come to process phones (i.e. vowels and consonants) in a manner specific to the language to which they are exposed [1]. For example, a baby exposed to English speech retain the ability to distinguish between the /r/ and /l/ sounds of English, while a baby exposed to Japanese, a language where there is no equivalent of this phonetic contrast, quickly learns to ignore it [2]. This phenomenon has been documented in a large number of studies but the mechanisms underlying it have been less studied. Because of the very early age at which it occurs -before the baby even speaks their first word- it has been proposed to result from some sort of statistical learning performed by the child on the basis of the speech signal reaching their senses. This invites further investigation into what specific input data and learning algorithm can plausibly account for the observed empirical results. A few proposals have been made [3–14], but the proposed models were never tested extensively nor compared quantitatively to see whether they are really able to account for a sizable portion of the available empirical observations.

The systematic comparison of models of phonetic learning is very important, from a theoretical perspective, but also for generating new empirical predictions that could be put to test in infants. This is why, we devoted this thesis, not to the modeling of phonetic learning itself, but to the preliminary step of developing a sound method to compare these models. To this effect, we introduce in this thesis ABX-discriminability measures, which provide a systematic and flexible way of evaluating candidate models for phonetic category acquisition. We demonstrate the interest of our evaluation framework, by applying it to the evaluation of models of phonetic category processing at birth and in adults. Models of phonetic category processing at birth provide an initial state for models of phonetic category acquisition. Models of phonetic category processing in adults provide a useful baseline against which to compare models of phonetic category acquisition, the difference between the two being that models of phonetic category processing in adults do not have to be based on a plausible learning mechanism, only the learning result needing to be plausible. The next step is to apply our framework to the models of phonetic category acquisition proposed in the literature, which is left for future work.

ABX-discriminability measures are useful beyond the particular problem of modeling phonetic category processing in humans and we also present other applications. There are at least

two ways in which the interest of ABX-discriminability measures generalizes to other situations. First, it generalizes to application domains beyond cognitive science. In particular, we discuss applications in artificial intelligence, low-resource engineering and data mining. Second, it generalizes to signal beyond speech and to category structures beyond phonetic categories. In this respect, although, we only work out practical examples involving large corpora of speech recordings annotated at the word or phone level, we present the rationale for applications in a fully general way.

The manuscript is organized in two parts. In Part I, we introduce ABX-discriminability measures and their applications. Chapter 1 defines ABX-discriminability measures, investigates their theoretical, statistical and computational properties and compare them to alternative methods of measuring the separation between categories. Chapter 2 presents three broad families of applications: evaluating systems operating with little or no explicit supervision in their ability to represent a category structure of interest; providing simple computational models of human or animal behavior in discrimination tasks; providing descriptive measurements for representations of categorical data. In Part II we study models of phonetic category processing at birth (Chapter 3) and in adults (Chapter 4).

Part I

Methods

Chapter 1

ABX discriminability measures: Theory

Contents

In this chapter, we introduce the notion of *ABX discriminability measures*. We study mathematical, statistical and computational properties of these measures that are useful in applications.

The basic intuition behind ABX discriminability measures, is that, given some tokens to which a category label is attributed (for example, some acoustic recordings of spoken words, with the associated phonetic transcriptions as category labels), we want to measure how well the different categories are separated in the token space by measuring whether tokens from the same category are closer to each other than to tokens from a different categories according to some measure of similarity between tokens. The A, B, X in ABX comes from the particular way this is done. Triplets A, B and X of tokens are formed where A and B are taken from different categories and X is either from the same category as A or as B. Then the similarity between A and X and between B and X is measured, and the triplets is rated as correctly classified if X is closer to the token from the same category. This is done systematically for all possible triplets from a given pair of categories and the proportion of correctly classified triplets is computed, yielding a measure of the separability of the two categories in the token space according to the similarity measure.

In most settings, the set of tokens considered is just a sample from all possible tokens of interest and we are not interested in learning something specific about the particular sample observed. The general device of probability theory allows taking this into account and we will present our results in a probabilistic framework.

In Section 1.1, we introduce ABX discriminability measures for the 2-category case. We then motivate their interest in Section 1.2, by comparing them to alternative methods of measuring the separation between two categories. In Section 1.3, we discuss ABX discriminability measures for an arbitrary number of categories. In Section 1.4, we discuss ABX discriminability measures for an arbitrary number of *structured categories*. By *structured categories*, we mean situations where each token is simultaneously characterized by several category labels (for example, some acoustic recordings of spoken words produced by different speakers, with the associated phonetic transcriptions *and* the speaker's identities as category labels).

1.1 ABX discriminability for two categories

In this section, we give a probabilistic definition of the notion of ABX-discriminability between two categories and show how it can be estimated in practice from a finite sample. More specifically, in Section 1.1.1, we introduce formally the probabilistic framework in which we work. In Section 1.1.2, we define the ABX-discriminability between two categories. In Section 1.1.3, we show how to estimate it from finite samples. In Section 1.1.4, we show how to compute confidence intervals around the estimated values.

1.1.1 Formalism and notations

Let us consider some observations $\mathbf{x} = (x_1, x_2, ..., x_m)$ and $\mathbf{y} = (y_1, y_2, ..., y_n)$ taking values in an arbitrary space E and associated to two different categories. For example, E could be a space of images and x could be a sample of images of animals and y a sample of images of inanimate objects. We consider the following probabilistic framework:

1. the sampling space E is supposed to be endowed with a σ-algebra Π, in order for (E, Π) to be a measurable space;

2. $\mathbf{x}$, respectively $\mathbf{y}$, is supposed to be an i.i.d. sample from probability measure $\mathbb{P}$, respectively $\mathbb{Q}$, on E and $\mathbf{x}$ and $\mathbf{y}$ are supposed to be independent of each other;

3. d is a measurable function from $(E \times E, \Pi \otimes \Pi)$ to $(\mathbb{R}, \mathcal{B})$ where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$. It is meant as a measure of dissimilarity between elements of E and we will refer to it as a *dissimilarity function*, but note that we do not require it to be a metric in the usual mathematical sense.

1.1.2 Definition and formal properties

1.1.2.1 Definition

Definition 1. The *ABX-discriminability* of a category with distribution $\mathbb{P}$ from a category with distribution $\mathbb{Q}$ according to dissimilarity function d is defined as the real number:

$$\mathcal{D}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q}) := p_{a,b,x\sim\mathbb{P}\otimes\mathbb{Q}\otimes\mathbb{P}}[d(a,x) < d(b,x)] + \frac{1}{2}\, p_{a,b,x\sim\mathbb{P}\otimes\mathbb{Q}\otimes\mathbb{P}}[d(a,x) = d(b,x)].$$

Remark 1. The second term in the addition above is important to ensure a consistent *chance level* (see Property 2 below). It is always equal to 0 and can be safely ignored when the dissimilarity function is continuous and the probability measures considered have no atoms.

Definition 2. The *symmetric ABX-discriminability* between $\mathbb{P}$ and $\mathbb{Q}$ according to d is defined as:

$$\Delta_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q}) := \frac{1}{2}(\mathcal{D}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q}) + \mathcal{D}_{\text{ABX}}(d, \mathbb{Q}, \mathbb{P})).$$

Definition 3. The *ABX-confusability* of $\mathbb{P}$ with $\mathbb{Q}$ according to d is defined as:

$$\mathcal{C}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q}) := 1 - \mathcal{D}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q}).$$

Definition 4. The *symmetric ABX-confusability* between $\mathbb{P}$ and $\mathbb{Q}$ according to d is defined as:

$$\epsilon_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q}) := 1 - \Delta_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q}).$$

Remark 2. The measurability of d ensures that all the quantities introduced in this section are well-defined.

1.1.2.2 Basic properties

In this section, we give some useful properties of ABX-discriminability measures. The proofs are straightforward and are not given.

Property 1 (Bounds)**.** *The quantities* $\mathcal{D}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q})$, $\Delta_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q})$, $\mathcal{C}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q})$, $\epsilon_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q})$ *defined above are comprised between* 0 *and* 1.

Property 2 (Chance-level)**.** *If* $\mathbb{P} = \mathbb{Q}$*, then:*

$$\mathcal{D}_{abx}(d, \mathbb{P}, \mathbb{Q}) = \Delta_{abx}(d, \mathbb{P}, \mathbb{Q}) = e_{abx}(d, \mathbb{P}, \mathbb{Q}) = \epsilon_{abx}(d, \mathbb{P}, \mathbb{Q}) = \frac{1}{2}.$$

Next, we define a function that is closely associated with ABX-discriminability measures and that we will encounter several times in this chapter.

Definition 5. Let us define:

$$\phi_d : \left| \begin{array}{ccc} E^3 & \longrightarrow & \mathbb{R} \\ a, b, x & \longmapsto & \mathbb{1}_{d(a,x)<d(b,x)} + \frac{1}{2}\mathbb{1}_{d(a,x)=d(b,x)} \end{array} \right. ,$$

where $\mathbb{1}$ denotes an indicator function, equal to 1 if its predicate is true and 0 otherwise.

The function we just defined finds its first usage in the following property, which provides an alternative characterization of ABX-discriminability measures that is often technically more convenient to manipulate than the original definition.

Property 3 (Alternative characterization)**.**

$$\mathcal{D}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q}) = \mathbb{E}_{a,b,x \sim \mathbb{P} \otimes \mathbb{Q} \otimes \mathbb{P}}[\phi_d(a, b, x)] = \int_{E^3} \phi_d(a, b, x) \mathbb{P} \otimes \mathbb{Q} \otimes \mathbb{P}(\mathrm{d}(a, b, x)).$$

1.1.2.3 Assymetry

In general, it is not true that $\mathcal{D}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q}) = \mathcal{D}_{\text{ABX}}(d, \mathbb{Q}, \mathbb{P})$, hence the interest of introducing the symmetric version $\Delta_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q})$. To illustrate how this asymmetry can be interpreted, let us study the quantity:

$$\rho_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q}) := \mathcal{D}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q}) - \mathcal{D}_{\text{ABX}}(d, \mathbb{Q}, \mathbb{P}).$$

Using property 3 together with the linearity of the integral, we obtain:

$$\rho_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q}) = \int_{E^2} \left[\int_E \phi_d(a, b, x) \mathbb{P}(\mathrm{d}x) - \int_E \phi_d(b, a, x) \mathbb{Q}(\mathrm{d}x) \right] \mathbb{P} \otimes \mathbb{Q}(\mathrm{d}(a, b)).$$

For any a, b elements of E, let us define $A_{ab} = \{x \in E \mid d(a, x) < d(b, x)\}$, $B_{ab} = \{x \in$

$E \mid d(a,x) > d(b,x)\}$ and $E_{ab} = \{x \in E \mid d(a,x) = d(b,x)\}$. For example, if $E = \mathbb{R}^n$ and d is the euclidean distance, A_{ab} is the half-space on the same side as a of the perpendicular bisector to segment $[ab]$ and B_{ab} the half-space on the other side. Then,

$$\phi_d(a,b,x) = \begin{cases} 1 & \text{if } x \in A_{ab} \\ 0 & \text{if } x \in B_{ab} \\ 1/2 & \text{else} \end{cases} .$$

Combining this with the observation that $\phi(a,b,x) = 1 - \phi(b,a,x)$ we obtain:

$$\rho_{\text{ABX}}(d,\mathbb{P},\mathbb{Q}) = \int_{E^2} \left[\mathbb{Q}(A_{ab}) - \mathbb{P}(B_{ab}) + \frac{1}{2}\left(\mathbb{Q}(E_{ab}) - \mathbb{P}(E_{ab})\right)\right] \mathbb{P}\otimes\mathbb{Q}(\mathrm{d}(a,b)).$$

Thus, ρ_{ABX} is obtained as an average over all pairs of points a and b, weighted by the probability distributions $\mathbb{P}$ and $\mathbb{Q}$ respectively, of the amount by which $\mathbb{Q}$ spills over the perpendicular bisector to segment $[ab]$ toward the a side, minus the amount by which $\mathbb{P}$ spills over that bisector toward the b side.

As an example, the support of two uniform distributions P and Q on the plane is represented in figure 1.1, P being rather *broad* and Q more *specific*. Taking two samples of distribution Q and one of distribution P, it's very likely that the two samples of distribution Q are going to be closer to one another than to the sample of distribution P. However, taking two samples of distribution P and one of distribution Q, it's much less likely that the two sample of distribution P are going to be closer to one another than to the sample of distribution Q. Thus in this example $\rho_{\text{ABX}}(d,P,Q) < 0$, i.e. P, the *broader* distribution is more likely to be confused with Q, the more *specific* distribution, than Q is likely to be confused with P.

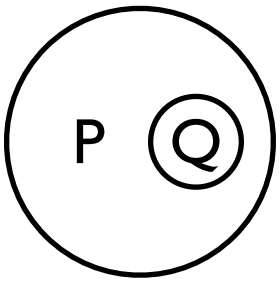


Figure 1.1: Two distributions illustrating the interpretation of asymmetry in the ABX-discriminabilities $\mathcal{D}_{\text{ABX}}(d,\mathbb{P},\mathbb{Q})$ and $\mathcal{D}_{\text{ABX}}(d,\mathbb{Q},\mathbb{P})$.

1.1.3 Point estimation

We gave a definition of ABX-discriminability measures in terms of probability distributions, but, in practice, we only have access to finite samples from these probability distributions. In Section 1.1.3.1, we introduce a finite sample estimator of the ABX-discriminability between two distributions. We show in Section 1.1.3.2, that this estimator has good statistical properties and is even optimal in a certain sense. In Section 1.1.3.3, we characterize the computational complexity of the proposed estimator. Finally, in Section 1.1.3.4, we introduced a sampled version of the estimator, allowing for trade-offs between computational complexity and statistical efficiency. The proofs for all the results are given in Appendix A.

1.1.3.1 A first estimator of the ABX-discriminability between two distributions

In practice, to estimate $\theta = \mathcal{D}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q})$, we usually only have access to independent i.i.d. samples $\mathbf{x}$ and $\mathbf{y}$ with law $\mathbb{P}$ and $\mathbb{Q}$ and size m and n respectively.

Definition 6 (Point estimator)**.** Let us define the *empirical ABX-discriminability* $\hat{\theta}$ as follows.

For any measurable function $d : E \times E \mapsto \mathbb{R}$ and any $\mathbf{x}$, $\mathbf{y}$ in $E^m \times E^n$,

$$\hat{\theta}(d, \mathbf{x}, \mathbf{y}) := \frac{1}{m(m-1)n} \sum_{a \in S(\mathbf{x})} \sum_{b \in S(\mathbf{y})} \sum_{x \in S(\mathbf{x}) \setminus \{a\}} \left(\mathbb{1}_{d(a,x) < d(b,x)} + \frac{1}{2} \mathbb{1}_{d(a,x) = d(b,x)} \right),$$

where $\mathbb{1}$ denotes an indicator function, $S(\mathbf{x})$ and $S(\mathbf{y})$ denote the multisets $\{x_1, x_2, ..., x_m\}$ and $\{y_1, y_2, ..., y_n\}$ constructed from the coordinates of $\mathbf{x}$ and $\mathbf{y}$ and $\setminus$ is the multiset difference operator.

1.1.3.2 Statistical properties

Let us now state some results showing that $\hat{\theta}$ is a good estimator of θ. Specifically, we show that $\hat{\theta}$ is *unbiased*, i.e. it does not systematically deviate from θ even for finite samples, *consistent*, i.e. it tends toward θ as the samples size increases, and *efficient* among unbiased estimators in a non-parametric setting, i.e. in the absence of additional assumptions on the probability distributions involved, $\hat{\theta}$ is on average closer to the true value of θ than any other unbiased estimator.

Property 4 (Unbiasedness)**.** *$\hat{\theta}$ is an unbiased estimator of θ.*

Property 5 (Consistency)**.** *$\hat{\theta}$ is a (strongly) consistent estimator of θ.*

$\hat{\theta}$ as an estimator of $\theta = \mathcal{D}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q})$ is optimal in the sense of the following theorem.

Theorem 1 (Efficiency)**.** *Among all estimators of $\mathcal{D}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q})$ that are unbiased for all possible probability distributions $\mathbb{P}$ and $\mathbb{Q}$ on (E, Π), $\hat{\theta}$ has minimum convex risk.*

1.1.3.3 Computational properties

In this discussion, to keep things simple, we ignore the case of ties. Taking it into account would not change significantly the computational complexity results. The following alternative expression for the empirical ABX-discriminability suggests a more efficient way to compute it than the original definition.

Property 6.

$$\hat{\theta}(d, \mathbf{x}, \mathbf{y}) = 1 - \frac{1}{m(m-1)n} \sum_{i=1}^{m} \sum_{j=1}^{m-1} (r_i(j) - j),$$

where $r_i(j)$ is the rank of the j-th element of $\mathbf{x}$ that is encountered when one goes through the elements of the multiset $\{d(e, x_i) |\ e \in (S(\mathbf{x}) \cup S(\mathbf{y})) \setminus \{x_i\}\}$ ordered from the smallest to the largest.

From the above formula, we can easily obtain a worst-case computational complexity for the computation of the empirical ABX-discriminability.

Property 7 (Computational complexity)**.** *$m(m-1)+mn$ evaluations of the dissimilarity function and $O(m(m+n)\log(m+n))$ elementary computing operations are always sufficient to compute $\hat{\theta}(d, \mathbf{x}, \mathbf{y})$.*

Property 7 shows that the worst-case computational complexity of our ABX-discriminability estimator is essentially quadratic in the size of the samples. For large samples, this can become problematic. In the next section, we show that the estimator can be easily adapted to offer a compromise between computational complexity and statistical efficiency.

1.1.3.4 Trading-off statistical efficiency for computational efficiency through a sampled estimator

A trade-off between computational complexity and statistical efficiency can be achieved by computing the ABX discriminability on only a subset of all possible ABX triplets. Here we consider selecting B triplets at random, with replacement, from the set of all possible triplets.

Definition 7 (Sampled point estimator)**.**

$$\hat{\theta}_B(d, \mathbf{x}, \mathbf{y}) := \frac{1}{B} \sum_{i=1}^{B} \left(\mathbb{1}_{d(a(i),x(i))<d(b(i),x(i))} + \frac{1}{2}\, \mathbb{1}_{d(a(i),x(i))=d(b(i),x(i))} \right),$$

where for any $1 \leq i \leq B$, $(a(i), b(i), x(i))$ is drawn at random with replacement from the multiset $\{(a, b, x) \in S(\mathbf{x}) \times S(\mathbf{y}) \times S(\mathbf{x}) \mid a \neq x\}$.

It is easily seen that $\hat{\theta}_B$ is still an unbiased and (strongly) consistent[1] estimator of θ. The next property, characterizes how the statistical efficiency of the estimator is affected by sampling.

Property 8 (Bound on the variance of the sampled estimator)**.**

$$Var[\hat{\theta}_B] \leq \frac{B-1}{B}\, Var[\hat{\theta}] + \frac{\theta(1-\theta)}{B} \leq \frac{B-1}{B}\, Var[\hat{\theta}] + \frac{1}{4B}.$$

The next property, characterizes how sampling affects the worst-case computational complexity of the estimator.

Property 9 (Computational complexity of the sampled estimator)**.** *It is sufficient to perform* $O(B)$ *elementary computing operations and to compute* $n(B) \leq m(m-1) + mn$ *dissimilarities to determine the numerical value of* $\hat{\theta}_B(d, \mathbf{x}, \mathbf{y})$.

The number of required evaluations of the dissimilarity function $n(B)$ is a random variable and it would be interesting to study its distribution further, in particular to understand the limits of the proposed sampled estimator in the cases where the computational bottleneck is in the calculation of the dissimilarities, but we leave this for future work.

[1] When $\min(m, n, B) \to +\infty$.

1.1.4 Interval estimation

In this section, we consider how to compute confidence intervals for ABX discriminability measures. We present a confidence interval based on a concentration inequality that is non-asymptotic but also quite conservative and a bootstrap confidence interval which is less conservative, but comes only with asymptotic theoretical guarantees. The proofs of the results are in Appendix A.

Theorem 2 (Non-asymptotic confidence interval)**.**

$$\left[\hat{\theta}(d,\mathbf{x},\mathbf{y}) - \sqrt{\frac{\log\frac{2}{\alpha}}{2\min\left(\lfloor\frac{m}{2}\rfloor,n\right)}};\ \hat{\theta}(d,\mathbf{x},\mathbf{y}) + \sqrt{\frac{\log\frac{2}{\alpha}}{2\min\left(\lfloor\frac{m}{2}\rfloor,n\right)}}\right]$$

is a confidence interval for θ with coverage at least α.

Let us define:

$$\rho_{01} = \frac{\mathrm{Cov}(\Phi_d(a_1,x_1,b),\Phi_d(a_2,x_2,b))}{\theta(1-\theta)}$$

and:

$$\rho_{10} = \frac{\mathrm{Cov}(\Phi_d(a,x_1,b_1),\Phi_d(a,x_2,b_2))}{\theta(1-\theta)},$$

where $\Phi_d : a,x,b \mapsto \frac{1}{2}\left(\phi_d(a,b,x) + \phi_d(x,b,a)\right)$.

Theorem 3 (Bootstrap asymptotic confidence interval)**.** *Suppose $\rho_{10} > 0$ and $\rho_{01} > 0$.*

Let α be the desired coverage of a confidence interval for θ. Let $\mathbf{X}^* = (X_1^*, X_2^*, ..., X_m^*)$ be i.i.d. random variables with common law $\frac{1}{m}\sum_{i=1}^m \delta_{x_i}$ and $\mathbf{Y}^* = (Y_1^*, Y_2^*, ..., Y_n^*)$ be i.i.d. random variables with common law $\frac{1}{n}\sum_{i=1}^m \delta_{y_i}$. Define $\hat{t}_{1-\frac{\alpha}{2}}$ and $\hat{t}_{\frac{\alpha}{2}}$ as the $1-\frac{\alpha}{2}$ and $\frac{\alpha}{2}$ quantiles of the distribution of

$$\sqrt{\min(m,n)}\left(\hat{\theta}(d,\mathbf{X}^*,\mathbf{Y}^*) - \hat{\theta}(d,\mathbf{x},\mathbf{y})\right).$$

If m and n tend towards $+\infty$ at a comparable rate, i.e. $\min(m,n) \to +\infty$ and there exists $\lambda \in]0;+\infty[$, such that $n/m \to \lambda$, then

$$\left[\hat{\theta} - \frac{\hat{t}_{1-\frac{\alpha}{2}}}{\sqrt{\min(m,n)}}, \hat{\theta} - \frac{\hat{t}_{\frac{\alpha}{2}}}{\sqrt{\min(m,n)}}\right]$$

is an asymptotic confidence interval for θ with coverage at least α.

1.2 Comparison with other measures of the separation of two categories

ABX discriminability measures, as defined in the previous section, provide a quantitative characterization of the separation between two categories. In this section, we compare them to other common ways in which the separation between two categories can be measured. In Section 1.2.1, we introduce several alternative measures. In Section 1.2.2, we investigate formally and empirically the relationship between ABX discriminability and these other measures. In Section 1.2.3, we discuss some elements that can help in deciding, for any given application, which method is the most appropriate. The proofs of the formal results in this section can be found in Appendix A.

1.2.1 Three types of measures of category separation

In this section, we introduce three (family of) alternatives to ABX discriminability measures for quantifying the separation between two categories: Same/Different (AX) discriminability, supervised classification and unsupervised classification. Same/Different (AX) discriminability measures the separation between two categories by asking to what extent it is possible to tell whether two tokens (A and X) come from the same category or not. As for ABX discriminability, this is based on measuring the performance in a discrimination task. Another way of characterizing the separation between two categories consists in measuring the performance in an X classification task, that is to say, measuring how difficult it is, given an unlabeled token from one of the category, to decide to which category it belongs. This can be done in (at least) two different ways. In *unsupervised classification*, also called *clustering*, the decision has to be made on the sole basis of an observed mixture of the distributions of the two categories, without knowledge of the individual distributions from which the mixture was generated. In *supervised classification*, by contrast, the distribution of each category is supposed to be available separately to inform the decision. We present Same/Different (AX) discriminability in Section 1.2.1.1, supervised classification in Section 1.2.1.2 and unsupervised classification in Section 1.2.1.3.

1.2.1.1 Same/Different (AX) discriminability

We begin by considering a measure of the separation of two categories that is also based on a discrimination task: Same/Different or AX discriminability. AX discriminability measures are based on a principle even simpler than ABX discriminability measures: given two tokens A and X, are they similar or different? In practice, as for ABX discriminability the decision is made on the basis of a measure of dissimilarity between A and X: if the dissimilarity is larger than a given threshold the two tokens are deemed different otherwise they are considered similar. For a given value of the threshold, the separability of the two categories can be characterized by two numbers: the *AX false positive rate* measuring the probability to consider the two tokens as different even though they were drawn from the same category and the *AX true positive rate* measuring the probability to consider the two tokens as different when they were really drawn from different categories. Formal definitions for these quantities are given below.

Definition 8. The *AX false positive rate* for a category with distribution $\mathbb{P}$ according to dissimilarity function d and threshold τ is defined as the real number:

$$\mathcal{F}_{\text{AX}}(d, \mathbb{P}, \tau) := p_{a,x\sim\mathbb{P}\otimes\mathbb{P}}[d(a,x) > \tau] + \frac{1}{2}\, p_{a,x\sim\mathbb{P}\otimes\mathbb{P}}[d(a,x) = \tau].$$

Definition 9. The *AX true positive rate* for discriminating a category with distribution $\mathbb{P}$ from a category with distribution $\mathbb{Q}$ according to dissimilarity function d and threshold τ is defined as the real number:

$$\mathcal{T}_{\text{AX}}(d, \mathbb{P}, \mathbb{Q}, \tau) := p_{a,x\sim\mathbb{Q}\otimes\mathbb{P}}[d(a,x) > \tau] + \frac{1}{2}\, p_{a,x\sim\mathbb{Q}\otimes\mathbb{P}}[d(a,x) = \tau].$$

It is often more convenient to summarize the separability of the two categories with a single number. Since there is no obvious way to set the value of the threshold τ, it is also desirable to find a measure that does not depend on the choice of a particular threshold τ. The Area Under the so-called ROC Curve, obtained when plotting the *AX true positive rate* as a function of the *AX false positive rate* as τ varies (see for example [15]), satisfy these requirements. We call this measure *AX discriminability*. It can be defined formally as follows, under some regularity conditions on the *AX false positive rate* and the *AX true positive rate* as functions of τ.

Definition 10. Let us suppose that $\tau \mapsto \mathcal{T}_{\text{AX}}(d, \mathbb{P}, \mathbb{Q}, \tau)$ is C^1 and $\tau \mapsto \mathcal{F}_{\text{AX}}(d, \mathbb{P}, \tau)$ is a C^1-diffeomorphism. Then the *AX discriminability* of a category with distribution $\mathbb{P}$ from a category with distribution $\mathbb{Q}$ according to dissimilarity function d is defined as the real number:

$$\mathcal{D}_{\text{AX}}(d, \mathbb{P}, \mathbb{Q}) := \int_{+\infty}^{-\infty} \mathcal{T}_{\text{AX}}(d, \mathbb{P}, \mathbb{Q}, \tau) \frac{\partial \mathcal{F}_{\text{AX}}}{\partial \tau}(d, \mathbb{P}, \tau) d\tau.$$

1.2.1.2 Supervised classification

Recall that our starting point is two independent i.i.d. samples $\mathbf{x} = (x_1, x_2, ..., x_m)$ and $\mathbf{y} = (y_1, y_2, ..., y_n)$ respectively from probability measure $\mathbb{P}$ and $\mathbb{Q}$ on a sampling space (E, Π). Our goal is to estimate from $\mathbf{x}$ and $\mathbf{y}$, without direct access to $\mathbb{P}$ and $\mathbb{Q}$, a measure $s(\mathbb{P}, \mathbb{Q})$ of the separability of $\mathbb{P}$ and $\mathbb{Q}$.

In a standard supervised classification setting, a set $\mathcal{H}$ of potential *classifier functions* is considered together with a real-valued *evaluation function* s that quantifies how well a given classifier can distinguish $\mathbb{P}$ from $\mathbb{Q}$. Then the idea is to take as an index of separability of $\mathbb{P}$ and $\mathbb{Q}$ the evaluation score s^* for the best classifier in $\mathcal{H}$, i.e.

$$s^*(\mathcal{H}, \mathbb{P}, \mathbb{Q}) := \sup_{h \in \mathcal{H}} s(h, \mathbb{P}, \mathbb{Q}).$$

Let us define more precisely the notion of classifier and evaluation function and give some examples. Roughly, a classifier is a function that takes as input a point of E and outputs a prediction regarding the category to which this point belongs. The prediction can be graded, more positive values indicating higher confidence that the point belongs to one category, and more negative values higher confidence that the point belongs to the other. More formally, a classifier can be defined as a (measurable) function from E to $\mathbb{R}$. For example, if $E = \mathbb{R}^p$, $\mathcal{H}$ could be the set of all linear classifiers, i.e. functions of the form:

$$h_{\mathbf{w},\mathbf{b}} : \left| \begin{array}{ccc} E & \longrightarrow & \mathbb{R} \\ & & \\ \mathbf{v} & \longmapsto & \mathbf{w}.\mathbf{v} + b \end{array} \right. ,$$

for $\mathbf{w}$ in E and b in $\mathbb{R}$ (where . denotes the canonical scalar product on E). The idea being that the equation $\mathbf{w}.\mathbf{v} + b = 0$ defines an hyperplane that separates E in two halves and that

points in one half will be attributed to one category and points in the other half to the other, with a confidence larger for points further away from the separation hyperplane. For example, Figure 1.2 shows two linearly separable probability distributions on $\mathbb{R}^2$ together with a separating hyperplane (an hyperplane in $\mathbb{R}^2$ being a line).

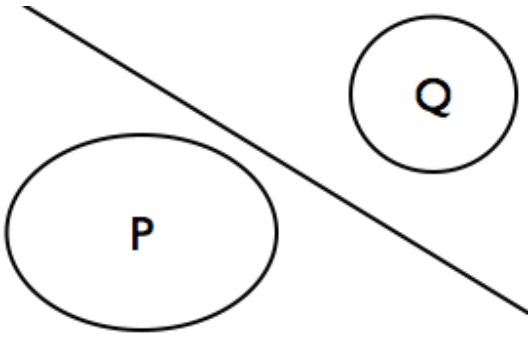


Figure 1.2: Two uniform probability distributions in the plane separated by a hyperplane.

Regarding the evaluation function, if we note $\mathcal{P}$ the set of all probability distributions on (E, Π), then the evaluation function s should simply be a function from $\mathcal{H} \times \mathcal{P} \times \mathcal{P}$ to $\mathbb{R}$. A simple example for s is the Area Under the ROC Curve (AUC) for binary classifiers:

$$s_{AUC}(h, \mathbb{P}, \mathbb{Q}) := p_{x,y\sim\mathbb{P}\otimes\mathbb{Q}}[h(x) < h(y)] + \frac{1}{2}\, p_{x,y\sim\mathbb{P}\otimes\mathbb{Q}}[h(x) = h(y)].$$

Of course, we cannot optimize directly s on $\mathcal{H}$ to find s^*, since we only have access to $\mathbb{P}$ and $\mathbb{Q}$ through the finite samples $\mathbf{x}$ and $\mathbf{y}$. Instead, we need a surrogate function $\hat{s}$ from $\mathcal{H} \times E^m \times E^n$ to $\mathbb{R}$, such that $\sup_{h\in\mathcal{H}} \hat{s}(h, \mathbb{P}, \mathbb{Q})$ is a statistically and computationally efficient estimator of $s^*(\mathcal{H}, \mathbb{P}, \mathbb{Q})$. There is a huge body of theoretical and practical work on the subject of finding an appropriate $\hat{s}$ for interesting choices of $\mathcal{H}$ and s. Efficient procedures with good theoretical guarantees have been found in a number of interesting settings (see for example [16] for a review).

1.2.1.3 Unsupervised classification

In an unsupervised classification or clustering setting, the starting point is a mixture $\mathbb{M}$ of $\mathbb{P}$ and $\mathbb{Q}$ with mixing weights w and $1-w$ respectively. The goal is to evaluate to what extent it is possible, on the basis of $\mathbb{M}$ alone, to find a finite partition of the input space E that separates $\mathbb{P}$ and $\mathbb{Q}$. The number of components to find (here 2) is sometimes but not always assumed to be

known. If it is assumed to be known, only partitions of the input space in 2 parts are considered.

The evaluation is performed in two steps. First, given a set of possible finite partitions $\mathcal{H}$ and an objective function $\mathcal{L}$, that rates the fit between $\mathbb{M}$ and any given partition, an optimal partition h^* is selected:

$$h^*(\mathcal{H}, \mathbb{M}) := \arg\min_{h \in \mathcal{H}} \mathcal{L}(h, \mathbb{M}).$$

Second, the ability of this partition to separate $\mathbb{P}$ and $\mathbb{Q}$ is rated by a scoring function s.

A finite partition can be represented by a function associating a given class to each point of E, i.e. a (measurable) function from E to $\{1, 2, ..., k\}$ for some k in $\mathbb{N}^*$.[2] For example, if $E = \mathbb{R}^p$, the set of possible partitions $\mathcal{H}$ could be the set of all k-means partitions, i.e. partitions defined by a set of k points of E, the centroids, such that the class of any point is that of the closest centroid. Formally:

$$h_{c_1,c_2,...,c_k} : \left| \begin{array}{ccc} E & \longrightarrow & \{1, 2, ..., k\} \\ v & \longmapsto & \arg\min\limits_{i \in \{1,2,...,k\}} ||v - c_i||_2 \end{array} \right. ,$$

for $(c_1, c_2, ..., c_k)$ in E^k. For example, a k-means partition of (a subset of) the plane, with $k = 20$ centroids is represented graphically in Figure 1.3.

The objective function $\mathcal{L}$ should be a function from $\mathcal{H} \times \mathcal{P}$ to $\mathbb{R}$, where $\mathcal{P}$ is the set of all probability distributions on (E, Π). A simple example is the k-means objective function, which measure the expected squared distance between a sample of $\mathbb{M}$ and the closest centroid:

$$\mathcal{L}_{km}(h_{c_1,c_2,...,c_k}, \mathbb{M}) := \mathbb{E}_{x \sim \mathbb{M}} ||x - c_{h(x)}||_2^2.$$

The scoring function s should be a function from $\mathcal{H} \times \mathcal{P} \times \mathcal{P}$ to $\mathbb{R}$. A simple example is the Rand index for equiprobable classes, which measures the probability that points that are really of different class are also in different parts of the partition and that points that are really of the

[2] Note that if the function h represents a partition and π is a permutation of $\{1, 2, ..., k\}$, then $\pi \circ h$ still represents the same partition. This means that in the proposed formalism s and $\mathcal{L}$ should be invariant to this kind of transformation applied to their first argument.

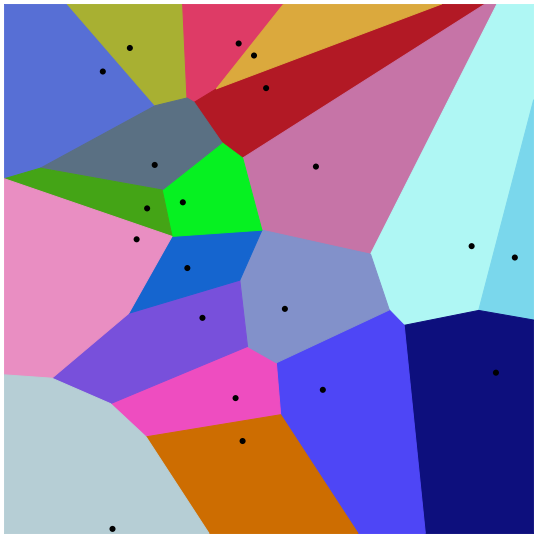


Figure 1.3: An example k-means partition with 20 centroids. Each colored cell represents the points belonging to the same class as the centroid represented as a black dot inside the cell (source: https://commons.wikimedia.org/wiki/File:Euclidean_Voronoi_diagram.svg).

same class are also in the same part of the partition:

$$\begin{aligned} 4s(h^*, \mathbb{P}, \mathbb{Q}) &:= p_{x,y\sim\mathbb{P}\otimes\mathbb{P}}[h^*(x) = h^*(y)] + p_{x,y\sim\mathbb{Q}\otimes\mathbb{Q}}[h^*(x) = h^*(y)] \\ &+ p_{x,y\sim\mathbb{P}\otimes\mathbb{Q}}[h^*(x) \neq h^*(y)] + p_{x,y\sim\mathbb{Q}\otimes\mathbb{P}}[h^*(x) \neq h^*(y)]. \end{aligned}$$

Of course, since we only have access to $\mathbb{P}$ and $\mathbb{Q}$ through the finite samples $\mathbf{x}$ and $\mathbf{y}$, we cannot optimize directly $\mathcal{L}$ on $\mathcal{H}$ to find h^* and we cannot evaluate directly s. Instead, we need a surrogate objective $\hat{\mathcal{L}}$ from $\mathcal{H} \times E^{m+n}$ to $\mathbb{R}$ and a surrogate score $\hat{s}$ from $\mathcal{H} \times E^m \times E^n$ to $\mathbb{R}$, such that:

$$\hat{s}\left(\arg\min_{h\in\mathcal{H}} \hat{\mathcal{L}}(h, \mathbf{z}), \mathbf{x}, \mathbf{y}\right)$$

is a computationally and statistically efficient estimator of:

$$s\left(\arg\min_{h\in\mathcal{H}} \mathcal{L}(h, \mathbb{M}), \mathbb{P}, \mathbb{Q}\right)$$

(where $\mathbf{z} = (x_1, x_2, ..., x_m, y_1, y_2, ..., y_n)$ is the concatenation of $\mathbf{x}$ and $\mathbf{y}$). Many choices of $\mathcal{H}$, $\hat{\mathcal{L}}$ and $\hat{s}$ are used in practice, but in most cases the available theoretical guarantees, when there is any, are much weaker than for supervised classification (see for example [17]).

1.2.2 Relationship with ABX discriminability

In this section, we investigate formally and empirically the relationship between ABX discriminability and the different types of evaluation methods presented in the previous section. In Section 1.2.2.1, we show that ABX discriminability and AX discriminability, although not identical, are formally connected and appear very closely associated in practice. We also show, in Section 1.2.2.2, that, although they are much looser, there are some formal connections and an empirical correlation between ABX discriminability and supervised and unsupervised classification measures.

1.2.2.1 ABX discriminability and AX discriminability

To relate AX and ABX discriminability our starting point is an interpretation of AX discriminability in terms of the expected ranking of the dissimilarities between a pair of similar and a pair of different tokens. This corresponds to a notion of ABXY discriminability, defined formally below.

Definition 11. The *ABXY discriminability* of a category with distribution $\mathbb{P}$ from a category with distribution $\mathbb{Q}$ according to dissimilarity function d is defined as the real number:

$$\mathcal{D}_{\text{ABXY}}(d, \mathbb{P}, \mathbb{Q}) := p_{a,b,x,y\sim\mathbb{P}\otimes\mathbb{Q}\otimes\mathbb{P}\otimes\mathbb{P}}[d(a,x) < d(b,y)] + \frac{1}{2}\, p_{a,b,x,y\sim\mathbb{P}\otimes\mathbb{Q}\otimes\mathbb{P}\otimes\mathbb{P}}[d(a,x) = d(b,y)].$$

Let us now state formally the correspondence between AX and ABXY discriminability.

Property 10. *Whenever* $\mathcal{D}_{\text{AX}}(d, \mathbb{P}, \mathbb{Q})$ *is defined:*

$$\mathcal{D}_{\text{ABXY}}(d, \mathbb{P}, \mathbb{Q}) = \mathcal{D}_{\text{AX}}(d, \mathbb{P}, \mathbb{Q}).$$

From the definition of ABXY discriminability, it is easy to characterize how it differs from ABX discriminability, which yields the following formal relationship between AX and ABX discriminability.

Theorem 4. *ABX discriminability and AX discriminability Let us define:*

$$\begin{aligned} p_1(d,\mathbb{P},\mathbb{Q}) := \quad & p_{a,b,x_1,x_2\sim\mathbb{P}\otimes\mathbb{Q}\otimes\mathbb{P}\otimes\mathbb{P}}[d(b,x_2) \leq d(a,x_1) < d(b,x_1)] \\ & + \tfrac{1}{2}\ p_{a,b,x_1,x_2\sim\mathbb{P}\otimes\mathbb{Q}\otimes\mathbb{P}\otimes\mathbb{P}}[d(b,x_2) \neq d(a,x_1) = d(b,x_1)], \end{aligned}$$

and:

$$\begin{aligned} p_2(d,\mathbb{P},\mathbb{Q}) := \quad & p_{a,b,x_1,x_2\sim\mathbb{P}\otimes\mathbb{Q}\otimes\mathbb{P}\otimes\mathbb{P}}[d(b,x_1) \leq d(a,x_1) < d(b,x_2)] \\ & + \tfrac{1}{2}\ p_{a,b,x_1,x_2\sim\mathbb{P}\otimes\mathbb{Q}\otimes\mathbb{P}\otimes\mathbb{P}}[d(b,x_1) \neq d(a,x_1) = d(b,x_2)]. \end{aligned}$$

Then, whenever $\mathcal{D}_{\mathrm{AX}}(d,\mathbb{P},\mathbb{Q})$ *is defined:*

$$\mathcal{D}_{\mathrm{ABX}}(d,\mathbb{P},\mathbb{Q}) - \mathcal{D}_{\mathrm{AX}}(d,\mathbb{P},\mathbb{Q}) = p_1(d,\mathbb{P},\mathbb{Q}) - p_2(d,\mathbb{P},\mathbb{Q}).$$

What is interesting is that although it is possible, it is not so easy to construct examples where p_1 and p_2 are very different from each other, suggesting that in many cases ABX discriminability and AX discriminability may be very close to each other numerically. It would be interesting to determine whether it is possible to bound theoretically the difference between p_1 and p_2, at least under some assumptions on d, $\mathbb{P}$ and $\mathbb{Q}$ that are likely to hold in practical applications. Given time constraints, we do not attempt to do this and instead, we perform an empirical study of the difference between ABX discriminability and AX discriminability in a practical example.

Contrast	**Preceding context**	**Following context**
ʌ-ɛ	s	n
ʌ-iː	t	n
ʌ-ɪ	SIL	n
ʌ-ɔɪ	p	n
d-ð	n	ʌ
k-n	ʌ	ʌ
k-t	ɪ	s
n-s	ʌ	t
n-s	ɛ	t
n-s	ɪ	t

Table 1.1: The 10 most frequent contrasts in a given phonetic context in a subset of the Wall Street Journal corpus (listed in no particular order). The symbol *SIL* is used when a segment is preceded or followed by silence.

Next, we explain the nature of the data in our practical example. The main point here is that it is a representative example of the kind of data to which we want to apply ABX discriminability, at least in the frame of this thesis. The practical details can be safely skipped, but we give them here for the interested reader. In this example, the tokens are phonetic segments as they occur in recordings of read speech. They are encoded as a sequence of MFC coefficients [18] (13 coefficients computed every 10 milliseconds). The number of coefficients characterizing a given segment thus depends on the length of this segment. As a notion of dissimilarity between such variable length items, we use Dynamic Time Warping (DTW) [19] based on a frame to frame cosine distance. A given token is characterized by the corresponding phonetic segment but also the identity of the preceding and of the following phonetic segments in the recording and the identity of the talker. The segments are extracted from recordings of 15 American English speakers reading news article from the *Wall Street Journal* (WSJ) corpus [20]. The time-boundaries of the segments are obtained by forced alignment using an HMM-GMM speech recognizer trained on the WSJ corpus. We consider the 10 most frequent pairs of phonetic segments in the corpus (listed in Table 1.1) and 15 speakers for which a large amount of data is available. For each speaker and each pair of segments, we compute ABX and AX discriminability scores. We chose frequent contrasts and speaker with a lot of data since our goal is to compare the theoretical quantity $\mathcal{D}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q})$ and $\mathcal{D}_{\text{AX}}(d, \mathbb{P}, \mathbb{Q})$ and not small sample estimates. The minimum number of tokens we obtain for each segment/talker pairs is 72 and we capped the maximum number of tokens as 200 to limit the duration of computation. The average number of tokens by segment/talker pair is 146.1.

The results are represented in Figure 1.4. The maximum absolute difference between the scores for the two methods is 13.2%. The (Pearson product-moment) correlation coefficient between the resulting $15 \times 10 \times 2 = 300$ ABX and AX discrimination scores is $r \approx .99$. (since ABX and AX discriminability are asymmetric, there are two scores for each talker/contrast pair). Although they are not identical, ABX and AX discrimination scores appear closely related in practice.

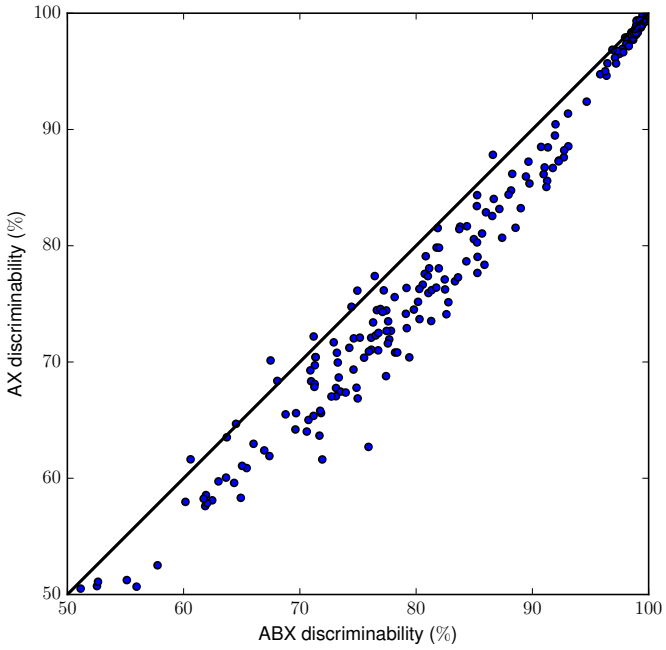


Figure 1.4: AX discriminability scores plotted as a function of corresponding ABX discriminability scores. The two measures are strongly correlated ($r \approx .99$). The line of equation $y = x$ is plotted in black.

1.2.2.2 ABX discriminability and X classification

ABX discriminability and X classification measures, supervised or unsupervised, are not as closely related as ABX and AX discriminability. However, some connections can still be established. In this section, we first formally show that a sample with perfect ABX discriminability can always be perfectly classified, in both a supervised and unsupervised manner. Then, we explore empirically the relationship between ABX discriminability and the performance of various classification algorithms.

Let us first introduce a simple supervised classification algorithm: k-nearest neighbors. In k-nearest neighbors a set of items with their associated category labels, the *training set* is provided. Given a new item to be classified, the k items from the training set that are closest to this item according to some distance function d are first retrieved. Then, the predicted category for the item to be classified is determined by a majority vote between the category labels associated to these k nearest neighbors. Let us suppose that a sample $\mathbf{z} = (z_1, z_2, ..., z_n) \in E^n$ from two categories as specified by category labels $\mathbf{c} = (c_1, c_2, ...c_n) \in \{1, 2\}^n$ is given. Let us also note $\mathbf{x}$, respectively $\mathbf{y}$, the multisets of elements from $\mathbf{z}$ from category 1, respectively category

2. Finally, let us assume that a partition (I, J) of {1, 2,..., n} into a training and test set is provided and let us note the training items and labels $\mathbf{z}_I = (z_i)_{i \in I}$ and $\mathbf{c}_I = (c_i)_{i \in I}$ and the test items and labels $\mathbf{z}_J = (z_j)_{j \in J}$ and $\mathbf{c}_J = (c_j)_{j \in J}$. We define the k-nearest neighbors classification accuracy $s_k(d, \mathbf{z}_I, \mathbf{c}_I, \mathbf{z}_J, \mathbf{c}_J)$ as the percentage of points from $\mathbf{z}^J$ that can be correctly classified (as specified by $\mathbf{c}_J$) using the k-nearest neighbor algorithm based on ($\mathbf{z}_I$, $\mathbf{c}_I$) and the distance function d.

The following theorem shows that if the empirical ABX discriminability of two categories is perfect, then it is always possible to perform supervised classification of these categories with perfect accuracy.

Theorem 5. *If the training set $\mathbf{z}_I$ contains at least one point of each class and*

$$\hat{\theta}(d, \mathbf{x}, \mathbf{y}) = 1 \text{ and } \hat{\theta}(d, \mathbf{y}, \mathbf{x}) = 1,$$

then the 1-nearest neighbor classification accuracy $s_1(d, \mathbf{z}_I, \mathbf{z}_J, \mathbf{c}_I, \mathbf{c}_J)$ is equal to 1.

Let us now state a similar theorem for unsupervised classification, using the same notations. We begin by introducing a simple unsupervised classification algorithm: single-linkage clustering. Single-linkage clustering is a form of agglomerative clustering in which each point is initially attributed to its own cluster. Then the clusters corresponding to the two closest points belonging to different clusters are merged. This merging step is then iterated until some stopping criterion is satisfied. Here we will only need the k-cluster stopping criterion ($k \leq n$), according to which the algorithm stops as soon as the number of clusters reaches k.

The following theorem states that if the empirical ABX discriminability of the two categories is perfect and the number of clusters to be found is known to be 2, then it is always possible to perfectly cluster $\mathbf{z}$ into the two categories.

Theorem 6. *If d is symmetric and:*

$$\hat{\theta}(d, \mathbf{x}, \mathbf{y}) = 1 \text{ and } \hat{\theta}(d, \mathbf{y}, \mathbf{x}) = 1,$$

then single-linkage agglomerative clustering on $\mathbf{z}$ with the 2-cluster stopping criterion yields the correct partition $\{\mathbf{x}, \mathbf{y}\}$.

These formal results only consider the case of perfect ABX discriminability. To get some sense of what happens when this is not the case, we perform an empirical comparison of ABX discriminability with various classification algorithms in the same practical example as in the previous section. For a fair comparison, we only use classification algorithm that can be applied to the same matrix of DTW dissimilarities between phonetic segments from which ABX and AX discriminability are computed.

We consider one supervised classification metric: k-nearest neighbor classification accuracy, as described earlier in this section, with two differences. First instead of using simple *hold-out* validation, we compute the scores using 5-fold cross-validation with random subsets chosen so that each subset contains the same proportion of points of each class as the initial sample. Second, when computing the accuracy, the same weight is given to confusions of items of class 1 with items of class 2 and confusions of items of class 2 with items of class 1, irrespective of the number of exemplars of each class in the sample. With this measure, the chance level is 1/2 as for ABX discriminability measures. For each computed score, k is selected among $\{1, 3, 5, 7, 9, 11, 13, 15\}$ to yield the best accuracy[3].

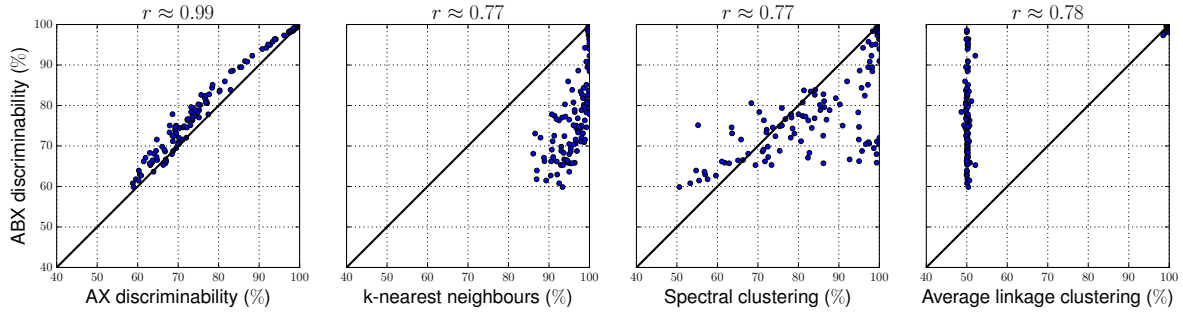


Figure 1.5: ABX discriminability scores plotted as a function of other evaluation metrics, with Pearson correlation coefficients reported above each plot. The line of equation $y = x$ is plotted in black.

There is a large variety of possible clustering algorithms and we consider two of them: average-linkage clustering with the 2-clusters stopping criterion (see for example [21], Section 14.3.12 pp. 523-526) and spectral clustering [22]. We chose average-linkage over single-linkage

[3]Note that the chance level after this selection is not strictly 1/2 anymore... To keep a chance level strictly at 1/2 and select k appropriately, a solution would be to split the initial sample into 3 parts, one as the training set, one to select k and one as the testing set. We ignore this here because we are considering fairly large samples (of size at least 72) and significant overfitting is unlikely to occur from the selection of k.

clustering because single-linkage, by merging clusters with minimal *minimal* distance between their points, tends to create clusters that are too elongated [23]. Average-linkage solves this problem by merging clusters with minimal *average* distance between their points. In spectral clustering, given a matrix of similarity between points, an objective function is defined over clusterings of size 2 measuring to what extent points within each cluster are similar to each other and dissimilar from points in the other cluster. A clustering minimizing a relaxation of this objective is then found using the spectrum of the similarity matrix, hence the name. In our case, we obtained the similarity matrix from the dissimilarity matrix by mapping each dissimilarity m in the matrix to $e^{-\frac{m}{s}}$ where s is the standard deviation between two elements of the dissimilarity matrix. The performance of both clustering algorithms is evaluated using the Rand index adjusted for chance [24], which yields a measure whose maximum is 1 and for which the chance level is 0. To simplify the comparison with the other measures, we rescale and shift the adjusted Rand index, so that the chance level is 1/2 while the maximum is still 1.

For simplicity we only consider symmetric versions of all the reported scores (e.g. for ABX discriminability we compute $\frac{1}{2}[\hat{\theta}(d, \mathbf{x}, \mathbf{y}) + \hat{\theta}(d, \mathbf{y}, \mathbf{x})]$). We thus obtain 15 speakers times 10 phonetic contrasts equals 150 scores for each method. The results are graphically represented in Figure 1.5. The main observation is that all three classification methods yields scores that are positively correlated with ABX discriminability scores, with a correlation coefficient around .77, independently of the method. Thus, while not as closely associated with ABX discriminability as AX discriminability, all classification measures still appear broadly consistent with ABX discriminability measures in practice.

Let us finish this section by looking at the correlations among the different classification measures in our empirical example, which yields another interesting result. The correlations between the scores obtained by the different methods are reported in Table 1.2. The most remarkable pattern in these correlations is that the classification measures are much less correlated with each other (last two columns of the table) than they are correlated with ABX- or AX-discriminability measures (first two columns of the table). This suggests that discriminability measures capture a fundamental aspect of the data of which the different classification measures can be seen as noisy variations.

ABX discriminability				
AX discriminability	.99			
k-nearest neighbors	.77	.75		
Spectral clustering	.77	.78	.66	
Average linkage clustering	.78	.83	.53	.56

Table 1.2: Pearson correlation coefficients between the scores obtained with different evaluation metrics. The metrics are ordered on the columns as they are ordered on the lines. Only the bottom-left half of the correlation matrix is represented because the correlations are symmetric.

1.2.3 Comparison of their properties as evaluation metrics

In the previous section, we observed empirically a broad similarity between the results obtained by different methods of measuring the separation between two categories. Given this similarity, how can we decide which method should be preferred in a particular application? Sometimes, the choice is obvious given the specifics of the application, but in many cases several alternatives appear possible (see Chapter 2 for concrete examples of applications). In this section, we discuss some general differences between the various methods that can provide some ground for deciding between them. In Section 1.2.3.1, we compare the scope of application of the different methods. In Section 1.2.3.2, we compare the free parameters in the different methods. In Section 1.2.3.3, we compare the computational complexity of the different methods. In Section 1.2.3.4, we compare the statistical efficiency of the different methods. Finally, in Section 1.2.3.5, we compare the methods with respect to a property that is required for certain classes of applications: *information neutrality*.

AX discriminability and ABX discriminability measures cannot be distinguished in terms of the properties we consider in this section and, for simplicity, we only discuss the case of ABX discriminability in the following. The reason for focusing on ABX discriminability measures in this thesis is that we expect them to be more easily related to measures of human or animal behavior (see Section 2.3.1) than AX discriminability measures, which would require either to manipulate in some way the participants decision threshold in an AX discrimination task or to have them perform a previously untested AXBY discrimination task.

1.2.3.1 Scope

We define the scope of application of a method as the minimal assumptions on its input that are required for the method to be usable in practice. Here we consider that for a method to be usable in practice it needs, at least, to be well-defined, computationally tractable and statistically consistent.

1.2.3.1.1 Definition Empirical ABX discriminability measures are well-defined given any dissimilarity function on the input space. For classification measures, the conditions of definition are not the same for all methods. Some methods, like k-nearest neighbor classification or spectral clustering are defined with the same level of generality as ABX discriminability measures. Others require more structure on the input space, like K-means clustering or SVM classification which require a vector space structure.

1.2.3.1.2 Computational tractability Here, we take *computationally tractable* to mean that a polynomial time algorithm to compute the measure is available. Empirical ABX discriminability measures are computationally tractable as long as the chosen dissimilarity function is computationally tractable. All classification methods used in practice are also computationally tractable, but many only yield approximative solutions to an underlying computationally intractable problem. For example, many supervised classification methods, like logistic regression and SVMs, replace the $0-1$ loss function with a convex loss to yield a tractable problem [25]. Unsupervised classification methods are also often tractable approximations to intractable problems, like spectral clustering which solves a convex relaxation of the intractable N-cuts problem [22] or K-means++ [26], which finds approximate solutions of the intractable K-means clustering problem.

1.2.3.1.3 Statistical consistency In most applications, results that would not generalize beyond the exact set of input data points available are not interesting. For example, if the input is a set of images of cats and dogs, the results are typically expected to apply to other sets of images of cats and dogs collected in a similar fashion. This means that having good statistical properties is an important requirement for any measure of the separation between two categories. In this section, we discuss the most basic statistical property one can ask from

an estimator: consistency.

We saw in Section 1.1, that, under the sole condition that the dissimilarity function d is measurable, the empirical ABX discriminability is a (strongly) consistent estimator of the ideal ABX discriminability for all possible underlying probability distribution. In comparison, classification measures have more limited guarantees. There is, to the best of our knowledge, no available theoretical guarantees when applying supervised classification algorithms with a dissimilarity function that is not a metric in the mathematical sense (in separable metric spaces, k-nearest neighbor classification, at least, is (weakly) universally consistent, under some regularity assumptions on the probability distribution of the points [27]). In the case of unsupervised classification, there is very little theoretical guarantees regarding the asymptotic behavior of most algorithms. Existing results include consistency of the ideal solution of K-means clustering (which is unfortunately computationally intractable), a weaker notion of consistency for single-linkage clustering and consistency of spectral clustering [17]. The consistency result for spectral clustering is proved in a quite general setting: for compact metric spaces with a similarity function that does not have to be related to the metric on the space and is only required to be symmetric, continuous and bounded away from zero by some constant.

1.2.3.1.4 Conclusion ABX discriminability measures have a wider scope than classification measures, all that is required being a measurable, computationally tractable dissimilarity function on the input space. In contrast, many classification methods require additional structure on the input space in their definition and/or are computationally intractable and require computationally tractable approximative solution to be used in practice. Some classification methods, such as k-nearest neighbors classification, are still both well-defined and computationally tractable under the same conditions as ABX discriminability measures, however available statistical consistency results for supervised classification algorithms are limited to the case of dissimilarity functions that are metric in the mathematical sense and for unsupervised classification algorithms are, in most cases, altogether absent, with the notable exception of spectral clustering. This is an important limitation of classification measures because non-metric dissimilarities are often used in practical applications. For example, in our numerical example in the previous section, we used the DTW dissimilarity, which does not verify the triangle inequality, to compare variable length representations. The Kullback-Leibler divergence [28] is another ex-

ample of a function that is not a metric and that is commonly used in practical applications, in order to compare probability distributions (we use it in Chapter 4 for instance). In contrast, statistical consistency for ABX discriminability measures holds for arbitrary measurable dissimilarity functions.

1.2.3.2 Free parameters

Free parameters in a procedure can sometimes be set on the basis of the specifics of the application at hand, which might suggest some obvious choices. If this is not the case, the parameters have either to be fitted to the input data, which complicates the procedure, or need to be set by hand, which complicates the interpretation of the results.

For example, the only free parameter of ABX discriminability measures is the dissimilarity function d. Fitting the dissimilarity function in full generality appears out of reach, so if the application does not suggest an obvious choice, some arbitrary default choice must be made in practice. If different reasonable default choices yield different results, the interpretation of the results become complicated. Even, the mere potentiality that a different reasonable default choice for it could yield different results complicates the interpretation of the results.

Yet, choosing a dissimilarity function for ABX discriminability is a comparatively weak requirement, when compared to setting the free parameters for classification methods. As we already saw, the free parameters of a classification method are a hypothesis space, an objective function and, for unsupervised classification, an additional scoring function. To see that setting a dissimilarity function is a lesser requirement, it suffices to notice that the definition of an objective function for supervised and unsupervised classification usually rely, implicitly or explicitly, on a notion of distance. For example, methods like k-nearest neighbors classification or K-means clustering rely explicitly on a notion of distance between input points. It is also the case for kernel methods, like kernel SVMs (see for example [29] Section 5.3), which rely only on dot products between input vectors and can be applied to any notion of distance between input points that can be expressed as a positive semi-definite similarity kernel.

Of course setting free parameters for classification methods implies committing to more than just a notion of distance: a hypothesis space and a learning algorithm need to be selected. There are many possibilities for these, often without any obvious default choice. In addition to that,

many learning algorithms rely on the fitting of some low-dimensional free parameters to the input data, which complicates both the practical implementation of the methods and the derivation of theoretical guarantees. Examples of such parameters include the number K of clusters in K-means clustering, the number of neighbors k in k-nearest neighbors or regularization parameters that control the bias/variance trade-off in structural risk minimization methods.

In conclusion, ABX discriminability measures have strictly less free parameters than classification methods, which simplifies their practical implementation and the interpretation of their results.

1.2.3.3 Computational complexity

We already discussed computational tractability, without which a method is essentially practically useless. Beyond computational tractability, lower computational complexity is generally desirable and can be very important in practice for applications with large amounts of data.

Method	**Time-complexity**
ABX discriminability	$\mathcal{O}(n^2 \log n)$
k-nearest neighbors	$\mathcal{O}(n^2 \log n)$
Average-linkage clustering	$\mathcal{O}(n^2 \log n)$
Spectral clustering	$\mathcal{O}(n^3)$

Table 1.3: Worst-case time-complexity for various measures of the separation between two categories given as a function of the total sample size n.

We compare the computational complexity for ABX discriminability measures and for some classification measures that are also defined and computationally tractable given arbitrary computationally tractable dissimilarity functions. Specifically, in Table 1.3, we compared worst-case time complexity for ABX discriminability measures and for the other methods considered in the numerical example of Section 1.2.2.2. The theoretical computational complexity is of the same order for ABX discriminability, k-nearest neighbors and average-linkage clustering. It is dominated by a term corresponding to sorting the pairwise dissimilarities between the n input points. The complexity for spectral clustering is a little worse and is dominated by a term corresponding to solving a generalized eigenvalue problem involving the matrix of pairwise similarities between the n input points. Note, however, that, for all four methods, in practice, the computation time is most often dominated by the computation of the n^2 pairwise dissimilarities. Also, note, that

although we are not aware of other classification methods applicable to arbitrary dissimilarity functions that would have a better computational complexity, certain methods with a more restricted scope can be faster, like Naive Bayes classification [30], for example, which is linear in the sample size.

In conclusion, ABX discriminability measures and classification methods with a comparable scope of application are essentially similar in terms of computational complexity.

1.2.3.4 Statistical efficiency

A classical way to characterize the statistical efficiency of estimators is to study their bias and their variance. ABX discriminability measures as we saw are unbiased and have minimum convex risk among all estimators of the ideal ABX discriminability that are unbiased for all possible underlying probability distributions. We also saw that they were concentrating around their mean with $\sqrt{n}$ precision, independently of the complexity of the input space or of the choice of the dissimilarity function. In contrast, there are no general unbiasedness or minimum risk guarantees for classification measures and results showing concentration around their means with a $\sqrt{n}$ precision can only be obtained if the hypothesis space considered has a limited complexity (see for example [31], Chapter 14). For unsupervised classification, there are very few theoretical guarantees on the bias and variance of the estimators, excepted for spectral clustering [17] with the same limitations as for supervised classification.

In addition, unsupervised classification methods are known to be very unstable in practice when the clusters are not well separated. Some approaches to characterizing theoretically this instability have been developed essentially in the particular case of K-means clustering (see mainly [32] and for a high-level review and discussion [33]), but it is observed in practice for all clustering techniques. Below, we show that already in a very simple example both K-means and spectral clustering suffers from this instability.

In this example, illustrated in Figure 1.6, the two classes have a uniform probability distribution over a segment of length L on the line and are separated by a distance d. We assume $L = 1$ without loss of generality by rescaling the axis. The interesting range of value for d then goes from -1 (where both classes are completely overlapping) to $+\infty$.

Through simulations, we obtained error rates as a function of d for two supervised classi-

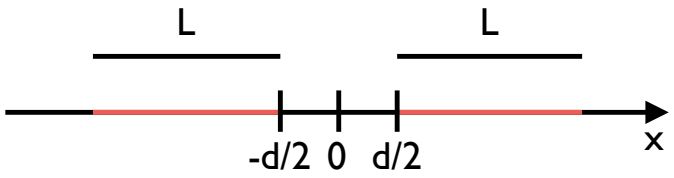


Figure 1.6: Two classes with uniform probability distributions over a segment of length L on the line, separated by a distance d.

fication methods - k-nearest neighbor classification and Linear Discriminant Analysis (LDA) - two unsupervised classification methods - K-means clustering and spectral clustering - and for ABX discriminability. We drew balanced random samples containing either 6 or 60 items in total (i.e. 3 or 30 of each class). For each of the two sample sizes, we drew 1000 samples for each of 100 values of d regularly spread between -1 and 3. We then applied each method to each sample. For k-nearest neighbor classification we used $k = 2$ neighbors[4] and for K-means, LDA and spectral clustering we fitted $K = 2$ classes. Error rates were obtained using the classification error for supervised classification methods and using the Rand index corrected for chance and rescaled between 0 and 1 for unsupervised classification methods (as in the numerical example of Section 1.2.2.2).

To study the variance and the bias of the different estimators, we also needed asymptotic error rates for the different methods in this simple example. For both supervised classification methods, it is easily seen that in this example as the sample size increase the error rate tends toward the Bayes classification error:

$$d \mapsto \begin{cases} -d/2 & \text{if } -1 \leq d < 0 \\ 0 & \text{if } d \geq 0 \end{cases}$$

For K-means clustering, it is possible to show through some tedious but elementary algebra that the asymptotic solution is formed of two clusters with centers symmetric around 0 (see Proposition 17), partitioning $\mathbb{R}$ into $]-\infty, 0[$ and $]0, +\infty[$ (0 can be randomly assigned to one or the other of the clusters without affecting the Rand index). Once again elementary algebra

[4]We use a k-nearest neighbor algorithm where the decision for classifying a new data point is taken by a majority vote among the k-nearest neighbors where each neighbor is given a weight inversely proportional to its distance to the data point to be classified. This is slightly different from what we did in Section 1.2.2.2 where each neighbor was given the same weight.

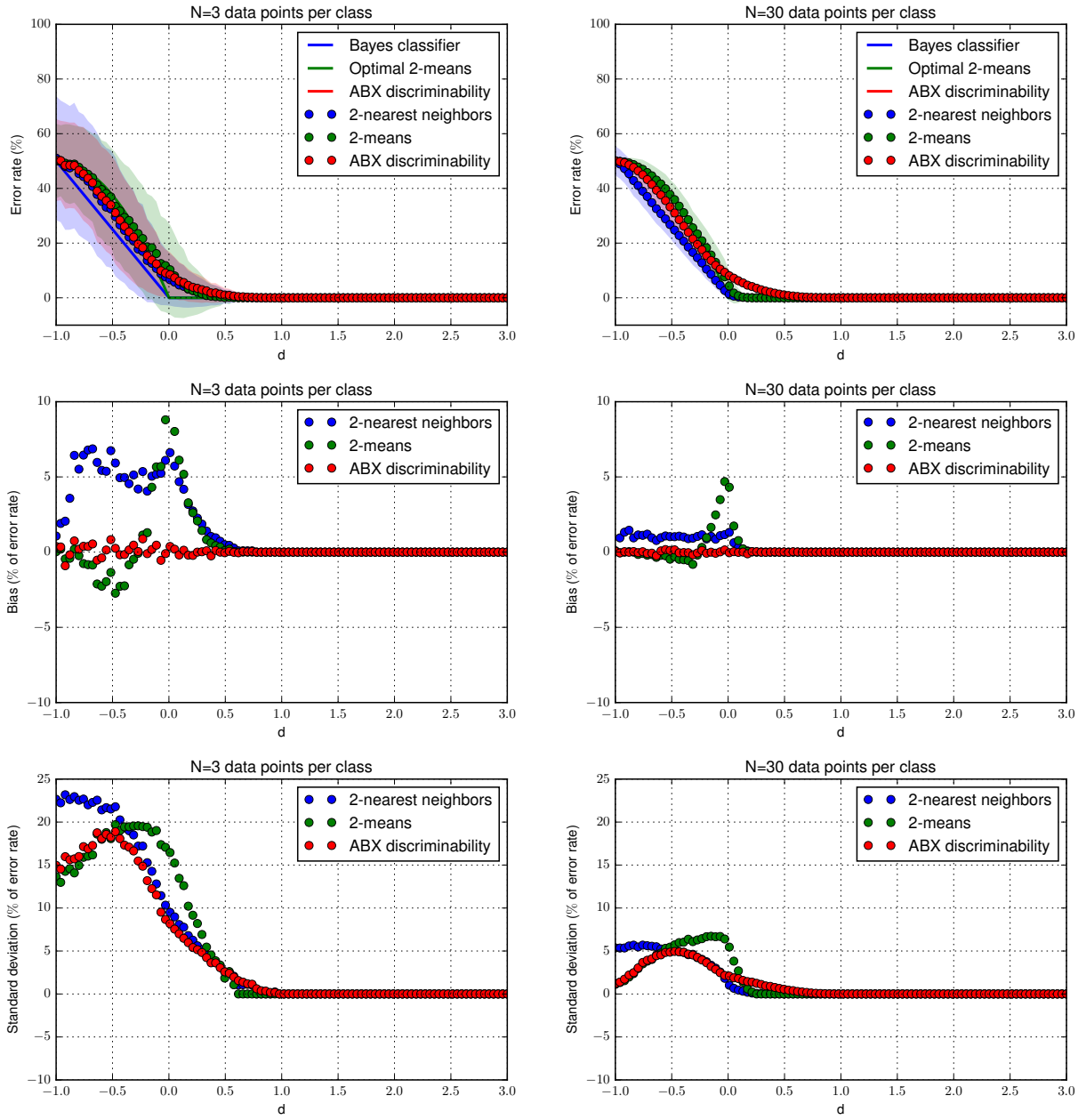


Figure 1.7: Comparison of the statistical efficiency of k-nearest neighbors classification, K-means clustering and ABX-discriminability in the line example. Results for balanced samples with 3 data points per class are reported on the left panels, results for 30 data points per class in the right panels. Top panels contain the asymptotic error rates as a function of d for each method (straight lines), the average value of the finite sample estimates for each method (dotted lines) and shaded confidence regions indicating the average of the finite samples estimates plus and minus their standard deviation. Middle panels contain the observed bias of the finite sample estimates as a function of d for each method. Bottom panels contain the observed standard deviation of the finite sample estimates as a function of d for each method.

shows that the rescaled and adjusted for chance Rand index for this particular partition is:

$$d \mapsto \begin{cases} \frac{1-(1+d)^2}{2} & \text{if } -1 \leq d < 0 \\ 0 & \text{if } d \geq 0 \end{cases}.$$

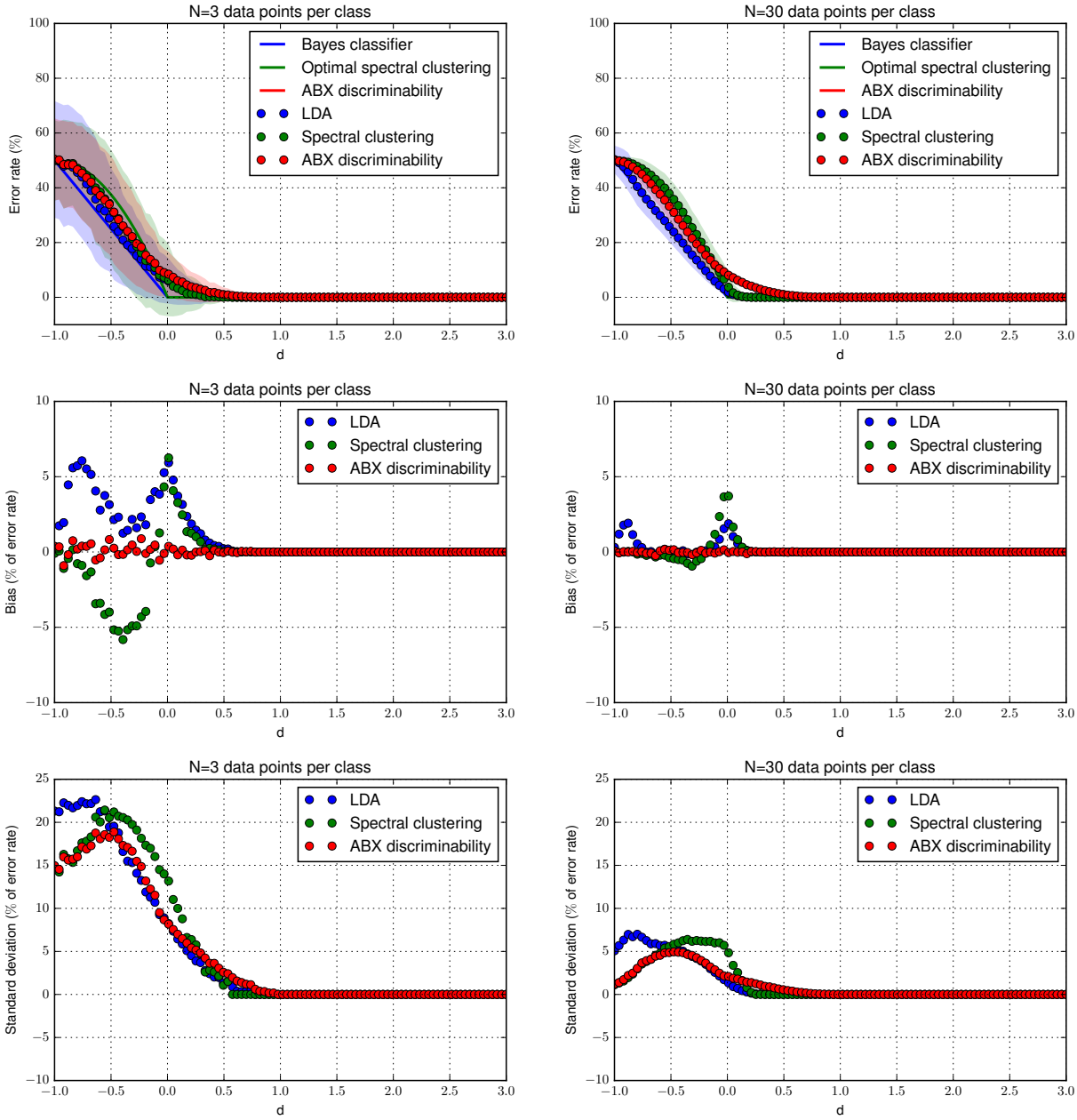


Figure 1.8: Comparison of the statistical efficiency of Linear Discriminant Analysis (LDA), spectral clustering and ABX-discriminability in the line example. Results for balanced samples with 3 data points per class are reported on the left panels, results for 30 data points per class in the right panels. Top panels contain the asymptotic error rates as a function of d for each method (straight lines), the average value of the finite sample estimates for each method (dots) and shaded confidence regions indicating the average of the finite samples estimates plus and minus their standard deviation. Middle panels contain the observed bias of the finite sample estimates as a function of d for each method. Bottom panels contain the observed standard deviation of the finite sample estimates as a function of d for each method.

For spectral clustering, we assumed that the asymptotic solution would also partition $\mathbb{R}$ into $]-\infty, 0[$ and $]0, +\infty[$ without seeking a formal proof. This seems intuitively right and predicts the same asymptotic Rand index as for K-means, which was consistent with our numerical simulations. Finally, for ABX-discriminability, the asymptotic error-rate is easily seen to be:

$$d \mapsto \begin{cases} \frac{1}{12}(7d^3 + 9d^2 - 3d + 1) & \text{if } -1 \leq d < 0 \\ \frac{1}{12}(1-d)^3 & \text{if } 0 \leq d < 1 \\ 0 & \text{if } d \geq 1 \end{cases} .$$

In Figure 1.7, we compare the statistical efficiency of k-nearest neighbors classification and K-means clustering with that of ABX-discriminability. We plot in the top panels the average observed error rate for each method for each of the two sample sizes together with confidence regions and the asymptotic error rates. In the middle panels, we plot for each method the observed bias (i.e. the difference between the expectation of the finite sample estimator and the asymptotic error rate) as a function of d. In the bottom panel, we plot the observed standard deviation of the finite sample estimators for each method. Figure 1.8 is organized as Figure 1.7, and contains the comparison of Linear Discriminant Analysis (LDA) and spectral clustering with ABX-discriminability.

As expected, the bias and variance of all estimators are reduced with larger samples. For a given sample size, simulations are consistent with the unbiasedness of ABX discriminability measures, while both supervised and unsupervised classification present some clear biases, in particular around $d = 0$ where their asymptotic error rate presents a kink which appears hard to estimate. In terms of variance of the estimators, supervised classification methods appear more variable than the other methods in the region $-1 \leq d < -.5$. Clustering methods appear more variable than the other methods roughly in the region $-.5 \leq d < .2$. ABX discriminability measures appear more variable than the other methods roughly in the region $.2 \leq d < .6$.

ABX discriminability measures appear better behaved overall, as they are unbiased and they are only more variable than the other methods in a region where the variability of all methods is already low. The higher variability of supervised classification methods when d is close to -1 is probably an artifact of using the classification error to measure error rate. Indeed, when d is close to -1, the two distributions highly overlap and the probability of

drawing finite samples where the sides of the two distributions appear the opposite of what they really are increases. Because the classification error, unlike the Rand index, is not invariant to permutation of the labels, it is much more affected by these events, leading to the observed increase in variability. Most importantly, the less reliable methods are clustering methods which, as expected, appear highly biased and unstable around $d = 0$. This nicely illustrates why it is difficult to characterize situations where two classes are neither completely overlapping nor completely separated with unsupervised classification methods. Unfortunately, these situations are often the most important in practical applications.

1.2.3.5 Information-neutrality

In this section, we define the property of *information-neutrality*, we explain its interest and, for each kind of evaluation methods considered so far, we look whether it has this property and under which conditions.

The various evaluation methods that we considered can all be broken down into two successive steps: first, a *task step* consisting in the realization of a task of interest on the basis of the input data and, second, a *grading step*, measuring how well this task was performed. We call a method *information-neutral* when the category labels are only required during the *grading step*.

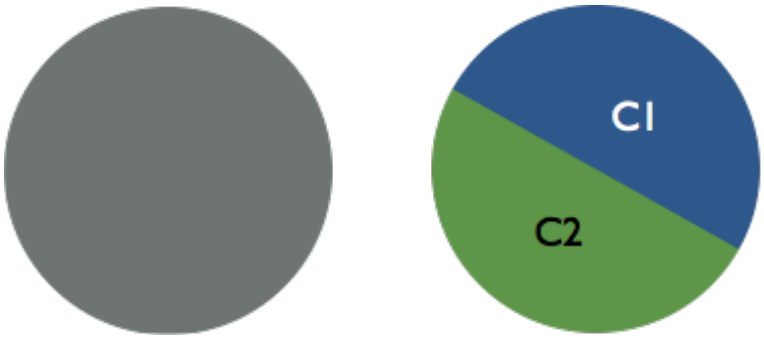


Figure 1.9: Two classes with uniform probability distributions over two opposite half-circles in the plane. Left: as they appear to a system that does not have access to the category labels. Right: as they appear to a system that can access the category labels. In this example, deciding whether a new point belongs to one or the other category (classification task) or deciding whether one point is of the same class or of a different class than other points (discrimination task) can be done perfectly if the category labels are available (because the probability distributions of the two classes do not overlap). If the category labels are not available it becomes completely impossible to do it[5].

[5]To make this statement precise, we need to not just consider the particular problem illustrated, which could

The practical importance of this concept is explained by the conjunction of two factors. First, there are whole classes of applications where the category labels cannot be assumed to be available during the *task step*. Second, assuming that the category labels are available during the *task step* can radically alter the nature of the evaluation task. Taken together, these two statements indicates that evaluation methods that are not *information-neutral* should not be used in applications where labels are not available during the *task step*. The first statement will be justified in the next chapter where we discuss practical applications (see Section 2.2.1). Let us explain now why the second statement is true. The difference in nature between tasks where the category labels are available (supervised learning) and tasks where they are not available (unsupervised learning) can be clearly seen by ignoring sampling issues (i.e. assuming that infinite training data is available). When category labels are provided along with the input data during the *task step*, one then has essentially direct access to the distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ of the individual classes as well as to their respective probability of occurrence in the training sample p and $1-p$, whereas, when category labels are not provided along with the training data during the *task step*, one only has access to the mixture distribution $\mathbb{P} = p\mathbb{P}_1 + (1-p)\mathbb{P}_2$. Because mixture distributions, in general, are not identifiable (i.e. many different combinations of probability distributions can result in the exact same mixture), the two situations are very different, as illustrated dramatically in Figure 1.9.

Let us now see which of the methods we considered are *information-neutral*. The natural breakdown when considering supervised classification as a measure of the separation of two categories, is to take the task of interest as that of attributing labels to new points given a set of labeled training examples. Then the grading step consists in measuring the agreement between the predicted labels and the actual labels. Since the task step in this breakdown of the procedure requires the knowledge of the labels of the points in the training set, supervised classification measures are not *information-neutral*. For unsupervised classification, the natural breakdown consists in taking as the task of interest the attribution of category labels to a set of unlabeled points and as the grading step, a measure of the agreement between the attributed labels and the actual labels for these points. The task step can be performed without any knowledge of the category labels of the input points, so that unsupervised classification measures

always be solved by a totally *ad hoc* procedure. Instead, we need to consider, for example, the whole class of problems with two uniform probability distributions over the two halves of a circle. Then it is easy to bound the possible performance of any unsupervised procedure *on average* over this whole class of problems.

are *information-neutral*. For ABX discriminability, the task of interest consists in taking triples of input points and asking whether the third point is closer to the first or the second. The grading step consists in measuring the agreement between the predicted answers for each triple and the answer according to the category labels. Thus, as long as the chosen dissimilarity function can be computed without any knowledge of the category labels of the input points, ABX discriminability measures are clearly *information-neutral*. Note that since an unambiguous answer can be derived from the category labels only if the first two points are from two different categories and the third point is from either of these categories, in practice only triples with this particular structure need to be considered. In summary, ABX discriminability and unsupervised classification are *information-neutral* but not supervised classification.

1.3 ABX discriminability for more than two categories

In this section, we extend the definition of ABX-discriminability to the case where there are more than two categories.

1.3.1 Formalism and notations

Let us suppose that we have some observations $\mathbf{x} = (x_1, x_2, ..., x_n)$ taking values in an arbitrary space E, with associated category labels $\mathbf{y} = (y_1, y_2, ..., y_n)$. We consider the following probabilistic framework:

- E and d verify the same properties as in Section 1.1;
- we assume, without loss of generality, that the label space is $\{1, 2, ..., K\}$ for some natural integer K;
- $\mathcal{D} = (x_i, y_i)_{i=1}^n$ is supposed to be an i.i.d. sample from some probability measure $\mathbb{P}$ on $(E \times \{1, 2, ..., K\}, \Pi \otimes 2^{\{1,2,...,K\}})$.

Let us consider a category $i \in \{1, 2, ..., K\}$.

Definition 12. The mixing weight p_i and probability distribution $\mathbb{P}_i$ (on E) associated to category i are defined as:

$$p_i := \mathbb{P}[E \times \{i\}]$$

and (assuming $p_i > 0$):

$$\forall X \in \Pi,\ \mathbb{P}_i[X] := \frac{\mathbb{P}[X \times \{i\}]}{p_i}.$$

Remark 3. $\mathbb{P}$ is completely determined by $(\mathbb{P}_i,\, p_i)_{i=1}^K$.

Notation 1. The number of observations with category label i is noted n_i and the corresponding observations are noted $\mathbf{o_i} = (o_{1,i}, o_{2,i}, ..., o_{n_i,i})$, for some arbitrary order of enumeration.

1.3.2 Definition

We propose to characterize the discriminability of K categories simply by the matrix of the pairwise discriminabilities.

Definition 13. The *ABX discriminability matrix* for categories with distribution $\mathbb{P}_1$, $\mathbb{P}_2$,..., $\mathbb{P}_K$ according to dissimilarity function d is defined as the real-valued matrix:

$$\mathcal{M}_{\text{ABX}}(d, \mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_K) := \begin{pmatrix} \mathcal{D}_{\text{ABX}}(d, \mathbb{P}_1, \mathbb{P}_1) & \mathcal{D}_{\text{ABX}}(d, \mathbb{P}_1, \mathbb{P}_2) & \ldots & \mathcal{D}_{\text{ABX}}(d, \mathbb{P}_1, \mathbb{P}_K) \\ \mathcal{D}_{\text{ABX}}(d, \mathbb{P}_2, \mathbb{P}_1) & \mathcal{D}_{\text{ABX}}(d, \mathbb{P}_2, \mathbb{P}_2) & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{D}_{\text{ABX}}(d, \mathbb{P}_K, \mathbb{P}_1) & \ldots & \ldots & \mathcal{D}_{\text{ABX}}(d, \mathbb{P}_K, \mathbb{P}_K) \end{pmatrix}.$$

The *ABX discriminability matrix* contains a lot of information and it is useful to also have lower-dimensional summary measures. A simple way to obtain summary measure is by linearly combining the pairwise discriminabilities.

Definition 14. Given a weight matrix $W = (w_{ij})_{1 \leq i \leq K,\ 1 \leq j \leq K}$ of real numbers, the *W-weighted average ABX discriminability* for categories with distribution $\mathbb{P}_1$, $\mathbb{P}_2$,..., $\mathbb{P}_K$ according to dissimilarity function d is defined as the real number:

$$\mathcal{D}_{\text{ABX}}(d, W, \mathbb{P}_1, \mathbb{P}_2, ..., \mathbb{P}_K) := \sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij} \mathcal{D}_{\text{ABX}}(d, \mathbb{P}_i, \mathbb{P}_j).$$

The most simple way to obtain a one-dimensional measure summarizing the discriminability of K categories, consists in taking the average pairwise discriminability, i.e. taking $w_{ij} = \frac{1}{K^2}$ for all i and j. As the discriminability of a category with itself is always $\frac{1}{2}$ and thus is not informative, it is usually not taken into account when computing the mean, i.e. the weights are

taken to be $w_{ij} = \frac{1}{K(K-1)}$ for all i and j in $\{1, 2, ..., K\}$ such that $i \neq j$ and $w_{ii} = 0$ for all i in $\{1, 2, ..., K\}$.

Many other weighting scheme can be useful. For example, in order to obtain an average discriminability specifically for category $k \in \{1, 2, ..., K\}$, we can take:

$$w_{ij} = \begin{cases} \frac{1}{2K} & \text{if} \quad i \neq j \text{ and } (i = k \text{ or } j = k) \\ 0 & \text{otherwise} \end{cases}.$$

1.3.3 Point estimation

Point estimators for the quantities defined in the previous section are obtained in a straightforward fashion from point estimators for the pairwise discriminabilities. The proof of the results in this section are trivial and are not given.

Property 11.

$$\hat{m}\,(d, \mathbf{x}, \mathbf{y}) := \begin{pmatrix} \frac{1}{2} & \hat{\theta}(d, \mathbf{o_1}, \mathbf{o_2}) & \dots & \hat{\theta}(d, \mathbf{o_1}, \mathbf{o_K}) \\ \hat{\theta}(d, \mathbf{o_2}, \mathbf{o_1}) & \frac{1}{2} & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\theta}(d, \mathbf{o_K}, \mathbf{o_1}) & \dots & \dots & \frac{1}{2} \end{pmatrix},$$

is a strongly consistent unbiased estimator for $\mathcal{M}_{\text{ABX}}(d, \mathbb{P}_1, \mathbb{P}_2, ..., \mathbb{P}_k)$.

Property 12.

$$\hat{\theta}\,(d, W, \mathbf{x}, \mathbf{y}) := \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbb{1}_{i \neq j} w_{ij} \hat{\theta}(d, \mathbf{o_i}, \mathbf{o_j}) + \sum_{i=1}^{K} \frac{w_{ii}}{2},$$

is a strongly consistent unbiased estimator for $\mathcal{D}_{\text{ABX}}(d, W, \mathbb{P}_1, \mathbb{P}_2, ..., \mathbb{P}_K)$.

1.4 ABX discriminability for structured categories

In this section, we consider the situation where the input data points are characterized simultaneously by several category structures. For example, if the input is speech, the signal corresponding to a given time-interval can be characterized by the sequence of phones uttered, but also by the identity of the talker (or talkers) or the topic of conversation. Each data point is thus described by a (fixed) number of category labels. We call this the case of *structured categories*, because

we can always see the tuple of category labels associated to each point as a single *structured* or *composite* category label.

The main interest of this *structured categories* case, is to allow the definition of richer ABX tasks which test to what extent the discriminability of the categories in a particular category structure is robust to the variability induced by one or several other category structures. For example, we can ask whether the discriminability of phones in a given representation of speech signal is robust to a change in talker.

1.4.1 Formalism and notations

Let us suppose that we have some observations $\mathbf{x} = (x_1, x_2, ..., x_n)$ taking values in an arbitrary space E, with associated category labels:

$$\mathbf{y_1} = (y_{1,1}, y_{2,1}, ..., y_{n,1}),\ \mathbf{y_2} = (y_{1,2}, y_{2,2}, ..., y_{n,2}),\ ...,\ \mathbf{y_c} = (y_{1,c}, y_{2,c}, ..., y_{n,c}),$$

for c different category structures. We consider the following probabilistic framework:

- E and d verify the same properties as in Section 1.1;
- we assume, without loss of generality, that the label spaces are $\{1, 2, ..., K_1\}$, $\{1, 2, ..., K_2\}$,..., $\{1, 2, ..., K_c\}$ for some natural integers K_1, K_2,...,K_c and we note $\mathcal{K}$ the cartesian product of these label spaces;
- $\mathcal{D} = \left(x_i, (y_{i,j})_{j=1}^{c}\right)_{i=1}^{n}$ is supposed to be an i.i.d. sample from some probability measure $\mathbb{P}$ on $(E \times \mathcal{K},\ \Pi \otimes 2^{\mathcal{K}})$.

Let us consider a (structured) category $\mathbf{i} = (i_1, i_2, ..., i_c) \in \mathcal{K}$.

Definition 15. The mixing weight $p_{\mathbf{i}}$ and probability distribution $\mathbb{P}_{\mathbf{i}}$ (on E) associated to category $\mathbf{i}$ are defined as:

$$p_{\mathbf{i}} := \mathbb{P}[E \times \{\mathbf{i}\}]$$

and (assuming $p_{\mathbf{i}} > 0$):

$$\forall X \in \Pi,\ \mathbb{P}_{\mathbf{i}}[X] := \frac{\mathbb{P}[X \times \{\mathbf{i}\}]}{p_{\mathbf{i}}}.$$

Remark 4. $\mathbb{P}$ is completely determined by $(\mathbb{P}_{\mathbf{i}},\ p_{\mathbf{i}})_{\mathbf{i} \in \mathcal{K}}$.

Notation 2. The number of observations with category label $\mathbf{i}$ is noted $n_{\mathbf{i}}$ and the corresponding observations are noted $\mathbf{o_i} = (o_{1,\mathbf{i}}, o_{2,\mathbf{i}}, ..., o_{n_{\mathbf{i}},\mathbf{i}})$, for some arbitrary order of enumeration.

1.4.2 ABX triples structure

ABX discriminability measures can be characterized by the structure of the ABX triples they specify. For example, for the measures we considered until now, i.e. measures with only one category structure, triples are chosen so that the A and X components are from the same category and the B component is from a different category, as illustrated in Table 1.4. In this case, we say that the measure is ON the category structure considered and we call the category common to A and X the ON_1 category and the category of B the ON_2 category. For example, we can perform a measure on the *phone* category structure associated to phonetic segments using, for instance, /b/ and /g/ as the ON_1 and ON_2 categories.

Category structure	A	B	X
1	ON_1	ON_2	ON_1
phone	/b/	/g/	/b/

Table 1.4: Structure of the ABX triples for ABX-discriminability measures based on only one category structure. The first line illustrates the general abstract pattern and the second line gives a specific example.

Now, given c different category structures, what structure for the triples of an ABX-discriminability measure are interesting ? In the most general case, the triples structure can be described as in Table 1.5, where no particular restriction is put on the choice of the c categories associated with each component of the ABX triple. Of course, not all possible structures are useful in practice and we restrict our discussion to certain configurations that provide easily interpretable measures. We explain in detail below which configurations we allow. They are summarized in Table 1.6.

Specifically, as in the case with only one category structure, we require the A and X components of a triple to be taken from the same category ON_1 of a given ON category structure and the B component to be taken from a different category ON_2 of this category structure. The interest of this condition is that it provides a clear answer to the question of whether X should

Category structure	A	B	X
1	α_1	β_1	χ_1
phone	/b/	/g/	/d/
2	α_2	β_2	χ_2
talker	Alice	Bob	Mary
$\vdots$	$\vdots$	$\vdots$	$\vdots$
c	α_c	β_c	χ_c
topic	Politics	Cinema	Science

Table 1.5: Structure, in the most general case, of the ABX triples for an ABX-discriminability measure based on multiple category structures. Every odd line illustrates the general abstract pattern for a given category structure and every even line gives a specific example.

be closer to A or to B. If it is not respected, i.e. all components are from different categories or all components are from the same category, there is no simple answer anymore.

One could imagine having a measure ON several category structures at the same time, however this would result in a measure confusing the impact of the different category structures on the discriminability. Instead, we recommend making distinct measures for each potential ON category structure of interest while keeping the other constants (see BY/WITHIN category structures below). So, we consider in the following that there is only one ON category structure and, without loss of generality, we assume that it is category structure 1.

What about the other category structures? The categories associated with the A, B and X components for a given category structure can either be:

- different for each component;
- the same for two components only;
- the same for all components.

If the categories are different for each component, they are different in particular for the A and B components. This comes in concurrence with the difference between those components introduced by the ON category structure, which complicates the interpretation of the measure, so we do not consider this case. If the categories are the same for two components only, for the same reason, it has to be for the A and B components. In this case, the category structure is called an ACROSS category structure, the category from which the A and B components are

taken is called the ACROSS$_1$ category and the category from which the X component is taken is called the ACROSS$_2$ category. A measure with an ACROSS category structure quantifies to what extent it is possible to detect that X is similar to A and different from B in terms of the ON category structure, despite X being different from both A and B in terms of the ACROSS category structure. There can be more that one ACROSS category structure of interest and in the following we assume that the ACROSS category structures are category structures $2, 3, ..., k+1$ for some k in $\{0, 1, ..., c-1\}$.

The only remaining possibility is to take the same categories for all components of the ABX triple for a given category structure. In this case, the category structure is called a BY/WITHIN category structure and the category from which all components are taken is called a BY/WITHIN category. A measure with a BY/WITHIN category structure quantifies to what extent it is possible to detect that X is similar to A and different from B in terms of the ON category structure, provided that A, B and X are all similar in terms of the BY/WITHIN category structure. There can be several BY/WITHIN category structures of interest and in the following we assume that the BY/WITHIN category structures are category structures $k+2, k+3, ..., k+l+1$ for some l in $\{0, 1, ..., c-1-k\}$. Remaining category structures that are not involved in the definition of the structure of the ABX triples can be completely ignored, i.e. we can assume without loss of generality that $k+l+1=c$.

1.4.3 Definition

Let us now define formally a notion of ABX-discriminability for structured categories. In this section, to keep things simple we assume that $k >= 1$ and $l >= 1$. The generalization to the case $k=0$ and/or $l=0$ is not complicated.

Notation 3. Let us note:

$$\mathcal{K}_o := \{1, 2, ..., K_1\}^2,$$

$$\mathcal{K}_a := \left\{ a_1, a_2 \mid a_1, a_2 \in \prod_{i=1}^{k} \{1, 2, .., K_{i+1}\} \text{ and } \forall i \in \{1, 2, .., k\},\ a_{1,i} \neq a_{2,i} \right\},$$

$$\mathcal{K}_b := \prod_{i=1}^{l} \{1, 2, ..., K_{k+1+i}\}.$$

Definition 16. Let us consider ON$_1$, ON$_2$ in $\mathcal{K}_o$, ACROSS$_1$, ACROSS$_2$ in $\mathcal{K}_a$ and BY in $\mathcal{K}_b$.

Role	Category structure	A	B	X
ON	1	ON_1	ON_2	ON_1
	phone	/b/	/g/	/b/
ACROSS	2	$\text{ACROSS}_{1,1}$	$\text{ACROSS}_{1,1}$	$\text{ACROSS}_{2,1}$
	talker	Bob	Bob	Alice
	⋮	⋮	⋮	⋮
	$k+1$	$\text{ACROSS}_{1,k}$	$\text{ACROSS}_{1,k}$	$\text{ACROSS}_{2,k}$
	preceding phone	/æ/	/æ/	/i/
BY/WITHIN	$k+2$	BY_1	BY_1	BY_1
	following phone	/i/	/i/	/i/
	⋮	⋮	⋮	⋮
	$k+l+1=c$	BY_l	BY_l	BY_l
	topic	Science	Science	Science

Table 1.6: General structure we consider for the ABX triples of ABX-discriminability measures based on multiple category structure. Every odd line illustrates the general abstract pattern for a given category structure and every even line gives a specific example.

The *ABX-discriminability* of category ON_1 from category ON_2 of category structure 1, across a change from categories ACROSS_1 to categories ACROSS_2 of category structures $2, 3, ..., k+1$, within categories BY of category structures $k+2, k+3, ..., c$, according to dissimilarity function d is defined as the real number:

$$\mathcal{D}_{\text{ABX}}(d, \mathbb{P}_a, \mathbb{P}_b, \mathbb{P}_x) := p_{a,b,x\sim\mathbb{P}_a\otimes\mathbb{P}_b\otimes\mathbb{P}_x}[d(a,x) < d(b,x)] + \frac{1}{2}\, p_{a,b,x\sim\mathbb{P}_a\otimes\mathbb{P}_b\otimes\mathbb{P}_x}[d(a,x) = d(b,x)],$$

where (with a slight abuse of notation):

$$\mathbb{P}_a := \mathbb{P}_{\text{ON}_1,\text{ACROSS}_1,\text{BY}};$$

$$\mathbb{P}_b := \mathbb{P}_{\text{ON}_2,\text{ACROSS}_1,\text{BY}};$$

$$\mathbb{P}_x := \mathbb{P}_{\text{ON}_1,\text{ACROSS}_2,\text{BY}}.$$

Remark 5. The notion of ABX-discriminability given above is asymmetric in the choice of ON_1 and ON_2 and in the choice of ACROSS_1 and ACROSS_2. Measures symmetrized for the choice of ON categories, the choice of ACROSS categories, or both, are easily obtained by permuting ON_1

and ON_2 and/or permuting ACROSS_1 and ACROSS_2 in the definitions of $\mathbb{P}_a$, $\mathbb{P}_b$ and $\mathbb{P}_x$ above and averaging the resulting measures.

The above definition is given for a specific choice of the ON_1, ON_2, ACROSS_1, ACROSS_2 and BY categories. Let us now define measure characterizing the pattern of ABX-discriminability more generally, for a specific choice of ON, ACROSS and BY category structures. As in Section 1.3, we define a matrix of ABX-discriminability between pairs of categories from the ON category structure.

Definition 17. The *ABX-discriminability matrix* ON category structure 1, ACROSS category structures $2, 3, ..., k+1$, BY/WITHIN category structures $k+2, k+3, ..., k+l+1$, according to dissimilarity function d, is defined as the K_1 times K_1 matrix $\mathcal{M}\left(d, (\mathbb{P}_{\mathbf{i}})_{\mathbf{i}\in\mathcal{K}}\right)$, with line i, column j element equal to:

$$m_{i,j} := \frac{1}{|\mathcal{K}_b||\mathcal{K}_a|}\sum_{b\in\mathcal{K}_b}\sum_{a_1,a_2\in\mathcal{K}_a}\mathcal{D}_{\text{ABX}}\left(d, \mathbb{P}_{i,a_1,b}, \mathbb{P}_{j,a_1,b}, \mathbb{P}_{i,a_2,b}\right).$$

Summary measures based on weighted sums of the elements in the ABX-discriminability matrix can be derived as explained in Section 1.3 and we do not describe it here again.

1.4.4 Point estimation

The proof of the results in this section are similar to those of the results from Section 1.1.3 or trivial and are not given. To define point estimators for ABX-discriminability measures based on structured categories, we first need to generalize our initial point estimator to the case where the distribution for A, B and X are all different.

Definition 18. Let us define the *empirical ABX-discriminability* $\hat{\theta}$ as follows.

For any measurable function $d : E \times E \to \mathbb{R}$ and $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ in $E^m \times E^n \times E^p$,

$$\hat{\theta}(d, \mathbf{x}, \mathbf{y}, \mathbf{z}) := \frac{1}{mnp}\sum_{a\in\mathbf{x}}\sum_{b\in\mathbf{y}}\sum_{x\in\mathbf{z}}\left(\mathbb{1}_{d(a,x)<d(b,x)} + \frac{1}{2}\,\mathbb{1}_{d(a,x)=d(b,x)}\right),$$

where $\mathbb{1}$ denotes an indicator function.

Property 13. *Let us consider i, j in $\mathcal{K}_o$, $i \neq j$, a_1, a_2 in $\mathcal{K}_a$ and b in $\mathcal{K}_b$. Then:*

$\hat{\theta}(d, \mathbf{o}_{i,a_1,b}, \mathbf{o}_{j,a_1,b}, \mathbf{o}_{i,a_2,b})$ is a strongly consistent, unbiased estimator of $\mathcal{D}_{\text{ABX}}\left(d, \mathbb{P}_{i,a_1,b}, \mathbb{P}_{j,a_1,b}, \mathbb{P}_{i,a_2,b}\right)$.

Property 14. *Let us consider i, j in $\mathcal{K}_o$, $i \neq j$. Then:*

$$\frac{1}{|\mathcal{K}_b||\mathcal{K}_a|} \sum_{b \in \mathcal{K}_b} \sum_{a_1, a_2 \in \mathcal{K}_a} \hat{\theta}(d, \mathbf{o}_{i,a_1,b}, \mathbf{o}_{j,a_1,b}, \mathbf{o}_{i,a_2,b})$$

is a strongly consistent, unbiased estimator of the line i, column j element of the ABX-discriminability matrix $\mathcal{M}\left(d, (\mathbb{P}_{\mathbf{i}})_{\mathbf{i} \in \mathcal{K}}\right)$.

Remark 6. We do not provide point estimators for the diagonal elements of the ABX-discriminability matrix because they have a slightly different form, which would require introducing additional notations, and they are not useful in practice.

Property 15. *Provided that the weights for diagonal elements are* 0 *(which is usually the case), strongly consistent, unbiased point estimators for summary measures based on weighted sums of the elements in the ABX-discriminability matrix are obtained by taking the corresponding weighted sum of the above estimators for the matrix elements.*

Chapter 2

ABX discriminability measures: Examples of application

Contents

In Chapter 1, we introduced ABX-discriminability measures and studied their properties from a formal point of view. In this chapter, we illustrate their practical interest in a variety of applications. In the context of this thesis, we only consider concrete examples where ABX-

discriminability measures are obtained from large corpora of speech recordings annotated at the word or phone level, but the reasons motivating our use of ABX-discriminability measures are not specific to this setting and apply just as well to other types of signals and/or category structures. Therefore, for each type of application considered, we motivate the use of ABX-discriminability independently of the particular nature of the signals and category structure considered. There still are some methodological aspects specific to the case where one computes ABX-discriminability measures from annotated corpora of speech recordings, which we discuss in Section 2.1. We then consider three broad classes of applications. In Section 2.2, we consider applications where the objective is to evaluate the performance of systems operating with little or no explicit supervision. We show that ABX-discriminability measures present significant advantages over existing alternatives when the objective is to evaluate such systems in relation to how they represent some category structure of interest. In Section 2.3, we consider applications where the objective is to computationally model discrimination tasks performed by humans or animals in behavioral experiments. We show how ABX-discriminability measures can be seen as simple abstract computational models for such tasks. Finally, in Section 2.4, we consider applications where the objective is to provide descriptive measurements characterizing datasets with category annotations.

Instances of the first class of applications -evaluating the performance of systems operating with little or no explicit supervision- are encountered in general when modeling how humans learn or when developing machines that can learn as efficiently as humans. As such this first class of applications is relevant, at least, to the fields of cognitive science, machine learning, artificial intelligence and low-resource engineering. The second class of applications -modeling discrimination tasks performed by human or animals in behavioral experiments- is obviously relevant to cognitive science in general and experimental psychology in particular. We discuss two specific instances from the third class of applications -providing descriptive measurements characterizing datasets with category annotations- but with the advent of the *big data* era, there is no doubt that this is only a tiny sample of the possible applications. The two examples we discuss are specifically: assessing whether phonetic categories are more distinct in infant-directed speech or adult-directed speech and measuring the impact of coarticulation on phone discriminability.

2.1 Methods: ABX-discriminability measures and annotated corpora of speech recordings

In Chapter 1, we introduced discriminability measures in an abstract setting, specifying neither the nature of the tokens to be discriminated nor that of the associated category labels. In this section, we focus on the specific case where the tokens to be discriminated are acoustic segments taken from some speech recordings and the associated category labels are derived from transcriptions of these recordings. We use the particular case of a lexical discrimination task as a guiding line in our discussion, as it proves sufficient to cover all the methodological points we need in the context of this thesis[1]. We discuss various possible approaches to the design of lexical discrimination tasks based on transcribed corpora of speech recordings. We first present the approach proposed by Carlin et al. [34] and discuss some of its limitations. Trying to overcome these limitations then leads us to the Minimal-Pair ABX (MP-ABX) discrimination tasks that we introduced in [35], whose benefits and drawbacks we also discuss.

A first approach was proposed by Carlin et al. [34]. In this approach, the discriminability of acoustic segments corresponding to whole words is computed in an AX discrimination task (see Section 1.2.1.1) based on stimuli from multiple speakers. More specifically, starting from a set of speech recordings with time-aligned annotations at the level of words, a representation of interest (e.g. MFCC coefficients or frame-level posteriorgrams from an acoustic model) is first derived for each word. Next, a measure of dissimilarity between the representations (e.g. a DTW dissimilarity based on the cosine distance or the Kullblack-Leibler divergence) is obtained for each possible pair of words and, given a dissimilarity threshold τ, two pairs are judged similar if their dissimilarity is less than τ and different otherwise. Finally, the overall discriminability is summarized independently of a particular choice for τ by either the *average precision* or the *precision at the precision/recall breakeven point*[2]. Different versions of these summary statistics can be obtained to characterize the discriminability of words either *irrespective of speaker identity*, *within speaker* or *across speaker* by computing the *recall* using respectively all pairs of words, only pairs of words uttered by the same speaker or only pair of words uttered by differ-

[1] Other types of tasks are of course possible. For example, we will also encounter some speaker discrimination tasks in the course of this thesis.

[2] Note that both of these measures are different from the AX discriminability measure that we defined in Section 1.2.1.1, which corresponds to the area under a curve defined by the *true positive rate* and the *false positive rate*.

ent speakers. In all three cases, even though this does not correspond to the usual definition, the authors propose to compute the *precision* as the proportion of same-talker same-word pairs present among all pairs that were judged similar irrespective of speaker identity.

Let us now discuss three limitations of this approach. A first important limitation of the approach proposed by Carlin et al. is that it does not allow to control the relative importance given to different word-types when computing summary statistics. Indeed, with this approach the influence of a given word-type on the precision and recall is essentially quadratic in the number of occurrences of that word-type in the corpus[3], even though in most applications there is no clear rationale for putting such an overemphasis over the most frequently occurring word-types. Since the relative frequencies of different word-types can vary a lot from one corpus to another (for example because of variation in speech register, format or topic), another unfortunate consequence is that the discriminability measures obtained are not appropriate to make comparisons across corpora. A second limitation of the approach proposed by Carlin et al. is that, while in some versions of the summary statistics the relative importance of the different speakers is controlled, other factors of variability are ignored. In particular, the phonetic context in which the words occur (i.e. the preceding and following words) is not controlled explicitly and it is easily seen, by analogy with the case of word-types, that this results in an essentially quadratic overemphasis on frequently occurring contexts. A third limitation of the approach proposed by

[3]This statement can be made precise, for example, by defining the *influence* of a given word-type on the precision or recall as the change in these quantities when one goes from a situation where the instances of that word-type are perfectly discriminable from instances of other word-types to the situation where they are completely confused with them, i.e. when all same-word pairs involving this word-type go from being judged similar to being judged dissimilar and all different-word pairs involving this word-type go from being judged dissimilar to being judged similar. In the following, we ignore speaker identity to keep things simple. Given a word-type w, let us look at its *influences* I_p and I_r on the precision and recall respectively, computed for an arbitrary threshold τ. We note n the number of occurrences of word-type w in the corpus considered, N the total number of occurrences of other word-types, S the number of same-word pairs in the corpus other than the same-word pairs of word-type w, F the number of pairs not involving word-type w that are judged similar at threshold τ and p the proportion of these F pairs that are actually same-word pairs. Then, we have:

$$I_r = \frac{n(n-1)/2}{S + n(n-1)/2},$$

and:

$$I_p = \frac{pF + n(n-1)/2}{F + n(n-1)/2} - \frac{pF}{F + nN}.$$

Under the assumptions that $n(n-1)/2 \ll S$, respectively $n(n-1)/2 \ll F$, we obtain:

$$I_r \approx \frac{n(n-1)}{2S},$$

respectively:

$$I_p \approx p + \frac{n(n-1)}{2F} - \frac{pF}{F + nN} \geq \frac{n(n-1)}{2F}.$$

Carlin et al. is that it ignores the phonological status of the words when forming different-words pairs. Since most words are phonologically very distinct (e.g. *car* versus *elephant*) and only words that are phonologically close are expected to be really hard to discriminate (e.g. *car* vs *cat*), the vast majority of the different-word pairs that are tested do not provide much information. This is not only a waste of computational power, but also a hindrance when computing summary statistics. Indeed, the choice of summary statistics by Carlin et al. appear rather *ad hoc* and statistics with better theoretical motivation, such as the Area Under the ROC Curve (AUC), that we proposed in Section 1.2.1.1 might seem preferable *a priori*[4]. However, because we are in this specific situation where most negative examples (i.e. different-word pairs) are of little interest, there is some evidence that the AUC might actually be less sensitive than a measure based on precision/recall curves like the *average precision* (see [36] for details).

Let us now consider ways of addressing the limitations we found with the approach proposed in [34]. First, we found that frequent word-types and phonetic contexts have a disproportionate influence on the proposed discriminability measures and more generally that variability factors other than speaker identity are not explicitly controlled. Explicit control can be obtained simply by conditioning on the value of the variability factors of interest when computing discriminability measures and aggregating the results with an appropriate weighting scheme. For example, if we want to control word-types and phonetic contexts, we need to compute summary statistic separately for each pair of word-type plus phonetic context, e.g. assessing the discriminability of *car* in the context of preceding *red* and following *parking* vs. *bit* in the context of preceding *a* and following *tired*. The other limitation we found, is that phonological status is ignored when forming different-word pairs, so that most pairs are too different to be of much interest. A simple way to address this limitation is to consider only pairs involving word-types that are minimally different phonologically, i.e. *minimal-pairs* of words differing only by one phone (e.g. *cat* vs. *kit*). An added benefit of considering only minimal-pairs is that a specific phonetic contrast can be attributed to each pair (e.g. the contrast /æ/-/ɪ/ for the pair *cat* vs. *kit*), allowing to derive a variety of fine-grained discriminability measures by aggregating the results appropriately. For example, we can obtain separate measures of the discriminability of vocalic contrasts, of contrasts involving only a change of place of articulation, of contrasts involving

[4]The AUC measures that we propose is better motivated theoretically in the sense that is has a simple interpretation as a ranking probability with a chance level of 50%. To the best of our knowledge, there is no such simple interpretation for the summary statistics proposed in [34].

only the suppression of a phoneme...

Combining conditioning on word-types and the restriction to minimal-pairs allows to overcome the main limitations of the approach proposed by Carlin et al., however this comes at a cost: minimal-pairs of words are not found that often in natural speech and even fewer will repeatedly occur within the same phonetic context. In other words, we obtain a theoretically more relevant measure but very large corpora of speech recordings might be necessary to reliably estimate it. There are several ways around this new problem. A first approach, consist in using speech corpora designed to include more minimal-pairs than is usual and in better controlled phonetic contexts. This is the approach we followed in [35] for example, where we used a manually time-aligned version [37] of the Articulation Index corpus containing isolated instances of all possible Consonant-Vowel stimuli in American English by 20 different speakers[5]. The main limit of this approach is that it requires specialized speech corpora that are not currently available for many languages. Another possible approach would be to try and estimate the discriminability of minimal-pairs from the discriminability of non-minimal pairs through some kind of regression model. The main limits of this approach are the added complexity and the fact that the reliability of the measure obtained will be ultimately limited by the reliability of the regression model chosen. A third approach, which we used in the Zero Resource Speech Challenge [38] and in various places in this thesis consists in forming minimal-pairs using segments that are not necessarily actual words. For example, one can consider minimal-pairs of syllables, single phones, diphones, triphones, etc. The main limit of this approach is that the pairs considered are not necessarily as ecologically relevant as minimal-pairs of actual words.

2.2 Application (i): Evaluating systems operating without explicit supervision

In the first class of practical applications we consider, the objective is to assess the performance of one or several systems in representing a category structure of interest, under the assumption that at least one of the systems considered operates without explicit supervision or with only a limited

[5]Of course we replaced the AX discrimination task of Carlin et al. by an ABX discrimination task. The main reason for this choice being that we expect ABX-discriminability measures to be more easily related to measures of human or animal behavior (see Section 2.3.1) than AX-discriminability measures, which would require either to manipulate in some way the participants decision threshold in an AX discrimination task or to have them perform a previously untested AXBY discrimination task.

amount of explicit supervision. In Section 2.2.1, we begin by explaining why this is a class of applications of significant practical interest. In Section 2.2.2, we argue that ABX-discriminability measures provide a better solution for this class of applications than more classic alternatives based on supervised or unsupervised classification. Finally, in Section 2.2.3, we review some results obtained in the specific case where the systems considered are speech processing systems and they are evaluated in their ability to support word or talker recognition.

2.2.1 A nonstandard but important evaluation context

We consider applications in which the objective is to assess the performance of one or several systems in representing a category structure of interest, under the assumption that at least one of the systems considered operates without explicit supervision or with only a limited amount of explicit supervision. In this section, we discuss why is this a practically interesting applicative scenario and introduce a specific example that is of particular relevance in the context of this thesis.

A major reason why the applicative scenario considered is of practical interest is that operating with no or little explicit supervision is commonplace for humans and animals, so that systems operating without explicit supervision are of interest at least in cognitive science (whose ultimate objective is to understand human cognition) and artificial intelligence (whose ultimate objective is to build machines that are as intelligent, or even more intelligent, than humans). Since human cognition involves the formation and manipulation of multiple category structures, the applicative scenario we considered appears relevant to both these fields. It is especially relevant for the study of early cognitive development. Indeed, in infancy the very limited communicative abilities of babies severely limit the amount of explicit supervision to which they can have access. In spite of this, infants have been shown to manipulate and learn about many category structures during their first year of life, such as phonetic categories [1], face identity, emotion or gender [39, 40], animated agents vs. inanimate objects [41–46], possible vs. impossible physical events [47–49], liquids vs. solids [50], number categories [51], etc. A second reason why the applicative scenario considered is of practical interest is that there is usually a non-negligible economic cost associated with the obtention of explicit supervision (which might explain why humans often operate without it in the first place). This suggests that the class of applications considered is also relevant to the field of low-resource engineering (which seeks

solutions to problems where the huge databases containing thousands or millions of explicitly annotated examples needed for training typical *supervised learning* systems are too costly or simply impossible to obtain). Finally, we can add that leading researchers in the *machine learning* field have identified the study of more challenging learning scenarios than the classic *supervised learning* setting (which includes for example the so-called *unsupervised*, *weakly supervised*, *semi-supervised* or *reinforcement* learning scenarios) as one of the most important and promising area for future developments (see for example the final sections in [52, 53]).

Let us now introduce a type of application of special interest in the context of this thesis: applications where the objective is to evaluate systems learning without explicit supervision from speech signal in a given language in their ability to represent words from this language. This is a type of problem relevant from the points of views of cognitive science, artificial intelligence and low-resource engineering at the same time. From the point of view of cognitive science, it is relevant when the speech processing systems considered are computational models of early language acquisition in infants. Indeed, being able to tell whether two speech sounds are two instances of a same word or instances of two different words, i.e. being able to categorize words on the basis of speech signal, is a fundamental ability that is necessary to enable verbal communication between human beings and that has to be learned, at least in the beginning, without explicit supervision. It has to be learned because word-forms and the basic sounds from which they are composed are not the same across languages. And empirical evidence that infants learn about the phonetic categories of their native language already during their first year of life [1], shows that it is acquired, at least in the beginning, in the absence of explicit supervision. By extension, the type of problem considered is relevant from the point of view of artificial intelligence as well, when the systems considered are artificial systems learning to process speech in the same conditions as humans. From the point of view of low-resource engineering, the type of problem considered is also relevant when designing speech processing systems for under-resourced languages. Indeed, being able to categorize words from the target language on the basis of speech signal is a fundamental requirement also in most practical applications in speech technologies. And for under-resourced languages there is no source of explicit supervision readily available, so that systems have to be trained without explicit supervision.

2.2.2 The benefits of ABX-discriminability measures in this context

So far, we introduced a class of applications and we explained why it was of significant practical interest in general as well as in the special case of evaluating speech processing systems in their ability to represent word categories. Next, we argue that ABX-discriminability measures provide a better solution for this type of applications than more classic alternatives based on supervised or unsupervised classification.

It might be surprising to propose evaluating systems in their ability to represent a category structure of interest based on measures of *discriminability*. Indeed the potential applications we outlined in the previous section might appear to involve *classification* tasks rather than *discrimination* tasks. For example, receiving a speech signal and identifying the words composing it seems more relevant to efficient verbal communication than receiving two different speech signals and deciding whether they correspond to the same word or not. Furthermore, if the goal is *classification*, can we not assume that the systems considered are classifiers, i.e. systems that map speech sounds to categorical labels ? This would have the benefit that the problem of evaluating classifiers has been studied for a long time and simple solutions are available such as the correct classification rate for supervised classifiers or the RAND index for unsupervised classifiers. In the remainder of this section we first explain why, in the applications we have in mind, it cannot be assumed that all the systems to be evaluated are classifiers. Then we argue that because we cannot make this assumption *and* because some of the systems we want to evaluate operate with little or no explicit supervision, *discriminability* measures are better for the applications considered than measures of *classifiability*.

There are at least two reasons why we cannot assume that all the systems to be evaluated are classifiers. The first is that, be it in cognitive science, artificial intelligence or low-resource engineering applications, explicit classification is usually not the ultimate objective of the system under study. The system does need to take into account certain category structures to produce an appropriate behavior or take appropriate decisions, but this does not necessarily imply that it has to represent these category structures explicitly. For example, when processing speech, human beings clearly take into account phonemic categories, in the sense that they treat acoustic word-forms that differ at least by one phoneme as linguistically similar and other word-forms

as linguistically different, independently from any semantic consideration. Yet, whether human speech processing involves an explicit representation of the signal as a sequence of phonemes has been the subject of a longstanding theoretical debate, which remains undecided. And even if humans did represent explicitly speech as a sequence of phonemes, the task involved would be more complex than a classification task, because it requires both to segment the signal into a sequence of segments and to classify these segments.

The second reason why we cannot assume that all the systems to be evaluated are classifiers is that, even if the ultimate objective was indeed explicit classification, the problems we are considering are often too complex to be tackled all at once. To make the discussion more concrete, let us explain this in the case of speech processing systems learning to represent word categories without explicit supervision. Because humans somehow do it, we know that it is possible in any language to learn to classify acoustic word-forms according to their lexical identity without explicit supervision. Yet our efforts to build artificial systems capable of the same achievement are only beginning, and state-of-the-art systems still have very poor performance [54]. Furthermore, while it is not plausible that infants have access to large amount of explicit supervision, there is still a lot of structure in the problem and many potential sources of weak supervision that infants could exploit (see for example [55–85]). To tackle problems of this level of complexity, one usually resorts to a reductionist approach, separating systems in smaller components and studying the behavior of these components in isolation or while they interact with each other. The practical result is that most of the time we want to evaluate partial solutions to the problem at hand. Since the partial solutions might involve many different formats of representation, even if the problem is one of explicit classification, we cannot assume that all the systems to be evaluated are classifiers. To give a concrete example, speech processing systems often take *speech features* as input, i.e. representations of speech that are not learned, such as spectrographic representation or MFC coefficients. There are many possible choices for these input representations and for the subsequent learning process, too many to be able to test all interesting combinations in practice. A simple heuristic to reduce the size of the search space is then to evaluate input representations independently of the subsequent learning process to select promising alternatives. Because the input representations can take a variety of different formats (discrete or continuous, sparse or dense, regularly sampled in time or not, etc.), we

cannot assume that all the systems to be evaluated are classifiers.

In summary, we are trying to evaluate systems, some of which operate without explicit supervision, in their ability to represent a category structure of interest and we do not want to make any *a priori* assumption regarding the format of representation used by these systems. Let us now show that ABX-discriminability measures offer a better solution to this problem than measures based on supervised or unsupervised classification. Perhaps the first idea that comes to mind to solve the problem considered would be to train a supervised classifier on the output of each system and compare the correct classification rate of the different classifiers. However, as we saw in Section 1.2.3.5, the presence of explicit supervision can completely change the nature and difficulty of a learning problem, so that the performance of a supervised classifier trained on the output of a system is not representative of the performance of that system if it has to operate without explicit supervision. To give a concrete example, if you are trying to develop a speech recognizer for an under-resourced language, knowing that if it had access to explicit supervision your system would be performant is not of much interest because you do not expect the system to ever have access to such explicit supervision. Thus in the applicative scenario we consider, evaluating systems based on supervised classifiers makes no sense.

To avoid the issue with supervised classification, one might naturally turn to solutions based on unsupervised classification (i.e. clustering) instead. There are some issues associated with evaluation based on clustering methods however. A first problem is to choose the classification method(s) to be used. There is no general solution to the problem of clustering and different methods are based on different assumptions and favor certain types of representations over other, independently of their intrinsic merits (for example diagonal-covariance Gaussian Mixture Models work better with de-correlated input representation). A possible solution is to treat the clustering method as a free parameter of the procedure, to be set independently for each representation that is tested. This is fairer than imposing the same method in all cases, but at the expense of the introduction of a complex free parameter in the evaluation procedure. A second issue is the lack of basic statistical guarantees for most clustering methods, especially when there are no strong assumptions on the format of the input representations (see Section 1.2.3.1.3 and Section 1.2.3.4). The computational complexity of clustering methods is also an issue as most clustering problems can only be solved approximately in practice, often in a

nondeterministic fashion. Finally, as we explained in Section 1.2.3.4, there appears to be an essential statistical instability of clustering methods that makes them particularly inappropriate for evaluation purposes. Indeed, there is a lot of statistical variability in clustering scores when the different categories to be clustered are close to being separated but not quite separated yet in the input representation, which is, unfortunately, a situation of great practical interest when evaluating and comparing different input representations.

Let us now explain why we think that ABX-discriminability measures offer a better alternative to measures based on supervised and unsupervised classification. First of all, as we explained in Section 1.2.3.5, as long as a dissimilarity function that can be computed without explicit supervision is used, ABX-discriminability measures, unlike measures based on supervised classification, characterize the performance of systems without assuming that explicit supervision is available[6]. Note that the claim is not that the proposed evaluation metrics are unsupervised themselves, as none of them can be computed without annotated examples[7], rather the claim is that they are appropriate to characterize how well systems operating without supervision will fare.

Thus far, we showed that, unlike measures based on supervised classification, ABX-discriminability measures are potentially appropriate for the problem considered. Let us now explain why we believe that they are in fact more appropriate than measures based on unsupervised classification. It might seem surprising to claim that *discriminability* measures better characterize systems learning to *categorize* than measures based on clustering. The rationale here is that, as we saw, measures based on clustering are very unstable in the regime that is the most interesting in practice (when problems are neither so hard that performance is close to chance nor so easy that they are perfectly solved) and that ABX-discriminability measures provide a much better-behaved surrogate. Indeed, ABX-discriminability measures are based on a very simple idea -that items of the same class should be close to each other and far from items from a different class- which underlie in one way or another all approaches to classification, yet they avoid the need to

[6]The crucial difference between, on the one hand, measures based on unsupervised classification and ABX-discriminability measures and, on the other hand, measures based on supervised classification, is that only the evaluation tasks associated with the former can be performed by the system on the basis of unlabeled exemplars (it is possible to decide whether stimuli X is closer to stimuli A or B without knowing the category label of either or even that X is either of the same category than A and of a different category than B or vice-versa).

[7]Indeed, category labels associated to the evaluation stimuli are still necessary in all cases in order to decide whether the answers given by the evaluated systems in the evaluation task are wrong or right and derive an average score.

actually commit to any particular clustering of the evaluation data. This allows to avoid both the intrinsic statistical instability associated with clustering (see Section 1.2.3.4) and the need to commit to a particular clustering algorithm (instead, one only needs to choose a dissimilarity function, which is a strictly more fundamental object and much more convenient to choose in practice, see Section 1.2.3.2) and yields measures with much nicer computational, statistical and theoretical properties (see Section 1.2.3). The pertinence of ABX-discriminability measures is corroborated by our empirical observations, in an example with speech stimuli (see Section 1.2.2.2), that measures obtained with different classification algorithms are more correlated with ABX-discriminability measures than they are correlated with each other, suggesting that ABX-discriminability measures capture some common ground of which the different classification measures are variations.

2.2.3 Examples of speech processing systems evaluated for word or talker recognition

To illustrate the potential of using discriminability measures for the evaluation of systems operating with little explicit supervision, we now review results obtained in the special case where the systems considered are speech processing systems and they are evaluated in their ability to represent word or talker identity.

Using performance in a discrimination task to evaluate speech processing systems in their ability to represent word categories was first proposed by Carlin, Thomas and Jansen in [34]. They proposed a Same/Different (AX) word-discrimination task, with performance measured either as the *mean average precision* (AP) or the *precision/recall at the breakeven point* (PRB) in this task (see Section 2.1 for more details). They looked separately at discrimination performance for words uttered by the same speaker and for words uttered by two different speakers only with the *breakeven point* measure. When computing the AP measure, they ignored talker identity, resulting in a measure that characterize mainly *across talker* discriminability[8]. They evaluated various systems on telephone speech recordings from the Switchboard American English and Fisher Spanish corpora and obtained a number of interesting results. For example, they compared discrimination scores obtained with various underlying dissimilarity functions

[8]This is because with a typical multi-talker speech corpus it is possible to form much more Same/Different pairs using words uttered by two different speakers than using words uttered by the same speaker.

and observed that among euclidean distance, cosine dissimilarity and symmetric KL-divergence, the cosine dissimilarity worked best for evaluating raw acoustic features and the KL-divergence worked best for evaluating representations taking the form of posteriorgrams. But perhaps their most important result, from the point of view of low-resource speech technology, is that applying directly to a low-resource language a system trained with explicit supervision on a high-resource language does not appear to work well. Indeed they found that discriminating words based on language-matched representations (best results 53.7% AP on Switchboard and 47.7% AP on Fisher) is substantially easier than discriminating words based on raw acoustic speech features (best results 21.5% AP on Switchboard and 10.6% AP on Fisher), whereas discriminating words based on language-mismatched representations (best results 16.2% AP on Switchboard and 10.3% AP on Fisher) is as hard or even harder than discriminating words based on raw acoustic speech features. Interestingly, through their within and across speaker PRB measures, they found that the improvement provided by language-matched representations over raw acoustic features was relatively modest when discriminating words within talkers and much bigger when discriminating words across talkers. This suggests that typical ASR systems trained with explicit supervision provide speech representations that are much more speaker-independent than raw acoustic features, but that this speaker-independence is highly language-specific. Since having speech representations that generalize across talkers is crucial in many ASR applications, this first study by Carlin, Thomas and Jansen established the need for innovative solutions for low-resource speech technology.

In following studies, only the AP measure of word discriminability was used, presumably because the focus was put on finding speech representations with good speaker-independence properties. In [86], it was assumed that clusters of similar word-sized segments of speech signal could be identified across speakers with a sufficient precision in the absence of explicit supervision through Spoken Term Discovery (STD) methods. This information was then used to train an HMM-GMM acoustic model[9]. AP scores obtained on the Switchboard corpus with the resulting acoustic model (29%) were higher than that obtained with raw acoustic features (16.9%) or with language-mismatched posteriorgrams from Neural Network (NN) systems trained with explicit supervision (16.7% and 8%), but remained largely below those obtained with language-matched posteriorgrams (51.6%). The AP measure was also used to study the dependence of the perfor-

[9]Without any other further supervision.

mance of the acoustic model on the value of some training parameters, including the number of sub-word units to be included in the acoustic model. The best performance was obtained for a number of sub-word units chosen around 100. In [87], an improved version of the acoustic model training procedure from the previous study [86] was proposed, demanding less from STD techniques while maintaining a comparable performance (28.6% AP instead of 29% AP). This new procedure consisted in training a GMM-based acoustic model without any supervision (with the standard Expectation-Maximization (EM) algorithm) and using information obtained from STD techniques to cluster the mixture elements into linguistically relevant groups. The authors also looked at the performance obtained directly with the GMM-based acoustic model without using the STD cues, which resulted in a more modest improvement over raw acoustic features (22.2% AP). In [88], the AP measure of word discriminability was used to assess the quality of speech features obtained from an unsupervised manifold learning procedure called *Intrinsic Spectral Analysis* (ISA). Because ISA features derivation is computationally expensive, the AP measure was computed on a smaller test set constituted from read speech taken from the TIMIT corpus of American English. The best ISA features yielded an AP score (48.5%) higher than raw features like MFC or PLP coefficients (33.8% and 34.8% respectively), but again much lower than language-matched posteriorgrams obtained from a system trained with explicit supervision (75.4%). on TIMIT, which I find very strange!!! The AP measure of word discriminability was also used in the context of a summer workshop organized by the CLSP at Johns Hopkins University in 2012 [54]. There it was shown that using raw acoustic features with coarser spectral resolution can improve the discriminability of words across speakers[10] (going from 17.7% AP to 21.2% AP on Switchboard for PLP coefficients for example). Also during this workshop, an approach to training aa HMM-GMM acoustic model without explicit supervision proposed in [90] was tested, but only on the TIMIT test set due to its computational complexity. It yielded an AP score of 44.5%, thus largely improving over raw acoustic features (best 34.8% AP) but remaining slightly inferior to ISA-based features (49.6% AP) and largely inferior to a supervised baseline (84.6% AP). Combining ISA features with the HMM-GMM training procedure from [90] did not provide further improvements (46.4% AP). Combining coarser spectral resolution with GMM-based acoustic model training with or without constraints from STD was not tested and none of these techniques were directly compared or combined with the unsupervised

[10] In accordance with the *front cavity* theory from [89].

HMM-GMM training procedure from [90] or ISA features extraction.

In [35], we introduced Minimal-Pair ABX discrimination tasks which provide a number of improvement over the approach proposed by Carlin et al. (see Section 2.1 for details). We used a systematic multi-speaker database of isolated Consonant-Vowels (CV) stimuli to implement various syllable discrimination tasks. While previous studies focused on *across speaker* discrimination, we also considered *across phonetic context* discrimination, phonetic context being another important factor of variability in the realization of acoustic word-forms. We also computed a complementary measure of talker discriminability across phonetic context. We applied our measures to the detailed evaluation of various steps in the derivation of the most common raw acoustic features in ASR. We found direct empirical support for many well-established principles in the design of speech features, such as the interest of using a non-linear scale of frequency or a compression of the dynamic range of the frequency channels to improve word discriminability, or the improved speaker-independence associated to coarser spectral resolutions, but we did not find clear-cut evidence of the benefits on word discriminability of casting features in the cepstral domain and/or of applying cepstral truncation or (temporal-domain) Linear Predictive Coding (LPC). For example, we found that casting MFC coefficients in the cepstral-domain slightly impairs the discriminability of words across talkers while simultaneously largely improving the discriminability of talkers across words[11]. In a follow-up paper [96], we extended our discussion to another important source of variability in the realization of acoustic word-forms: the presence of additive or convolution noise. Once again the benefits of spectral frame modeling by cepstral truncation or (temporal-domain) LPC were not obvious. We found, however, a clear benefit of RASTA filtering of the frequency channels in the presence of convolutional noise and of Frequency-Domain LPC (FDLP) in the presence of both additive and convolutional noise. The results of these two studies are summarized and discussed in the context of modeling speech processing by humans at birth in Chapter 3 of this thesis.

[11] Although this is not our focus in this section, discrimination measures have also been proposed to be of potential interest for purely supervised system design [34, 54, 91] as a rapid evaluation method that does not commit to a particular downstream processing architecture. Our results showing that casting MFCC in the cepstral domain improve talker discriminability but not phone discriminability can be interpreted in this context. Indeed, they are in line with recent development in speech features design associated with the replacement of GMM-HMM-based technology by NN-based technology: state-of-the-art ASR systems are now based on Mel-spectrograms (see for example [92, 93]), while state-of-the-art speaker recognition systems are based on MFC coefficients (see for example [94]). See also [95] for a concrete example of using ABX-discriminability measures to compare the coding properties of magnitude-spectrum-based and phase-spectrum-based speech features and translating the results into improvements in a standard (supervised) talker recognition application.

Beyond raw acoustic features, ABX-discriminability measures have also been applied to the evaluation of systems *learning* from speech signal without explicit supervision. In [97], as in previous papers that we already discussed [86, 87], the potential of using information obtained through STD techniques to train an acoustic model without supervision was investigated. As in the previous studies, the focus was on training an acoustic model and an idealized version of the information about similar chunks of signal supposed to be provided by STD techniques was extracted from manual annotations of speech recordings. While [86, 87] considered HMM-GMM and pure GMM acoustic models, this new study trained an NN-based acoustic model using a Siamese Neural Networks learning procedure to exploit the available similarity information. The system was trained and tested on different subsets of the TIMIT database. In a cross-talker ABX-discriminability task, the performance of the system (11.8% error rate) was largely above raw acoustic features (19.5% error rate). It was actually better than the performance of an HMM-GMM supervised ASR system trained with HTK on the same dataset (11.0% error rate) and close to the performance of a similar neural network trained with explicit supervision (9.2% error rate). To consolidate and promote synergies between the burgeoning efforts in the low-resource speech technology and in the infant language acquisition research communities, we also organized the Zerospeech challenge [38, 98], which was featured as a special session at the Interspeech 2015 conference. The challenge aim was to provide theoretically motivated, well-documented and freely available evaluation metrics and datasets that would allow to assess the state-of-the-art in the development of speech processing systems learning without explicit supervision. The challenge contained two tracks, the first of which was dedicated to the evaluation of speech representations through minimal-pair ABX tasks. The evaluation was carried out using recordings from two freely accessible corpora of continuous speech: the Buckeye Corpus of American English [99] and the NCHLT corpus of Xitsonga [100]. For each corpus, we implemented a within-talker and an across-talker ABX discrimination task based on minimal-pairs of triphones differing only in their central phone. Results of the initial session of the challenge have been summarized in [98] and the challenge remains open online, with all the evaluation code and resources freely available. In the following we briefly review the results of the challenge and for the sake of concision unless explicitly stated otherwise, we only report error rates in the cross-speaker task on the Buckeye corpus (which are quite representative). Solutions exploit-

ing cross-speaker word-level similarities discovered through STD methods to train an acoustic model were tested in the challenge, based for the first time on information actually obtained through STD methods (rather than using an idealized version obtained from manual annotations of speech recordings). The Siamese Networks approach from [97] was tested [101] and compared to an approach [102] also training an NN-based acoustic model but using Correspondence Autoencoders. Both approaches fared much better than raw acoustic features (28.1% ABX error rate) and close to a supervised HMM-GMM baseline (16.0%), with the Siamese Network approach (17.9%) having an edge over the Correspondence Autoencoders approach (21.1%) and even beating the supervised baseline in one case[12]. Methods using STD-derived information to train GMM or HMM-GMM acoustic models were not tested in the challenge, but later results from [91] on the earlier AP evaluation metric suggest that they are not as performant[13]. A study published after the initial session of the challenge [103] also showed that replacing the original input features (Mel-filterbanks) with richer scattering spectrum coefficients [104] improved the performance of Siamese Network acoustic models (17.1% ABX error rate). Other participants to the challenge proposed to model directly the speech signal at the frame level, without using word-level similarities obtained by STD methods. Badino et al. [105] trained unsupervised auto-encoders, yielding results slightly improving over the raw acoustic features baseline (26.3% ABX error rate). Chen et al. [106] trained a DPGMM-model working on frames of standard acoustic speech features without any external source of supervision, yielding surprisingly good results (16.3% ABX error). They actually obtained the best results in the challenge on all tasks except for the across speaker ABX task on the NCHLT corpus where the Siamese Network had the edge (16.6% ABX error rate for the Siamese Network vs. 17.2% ABX error rate for the GMM). These results are particularly surprising when compared to the earlier results from [87] that we already mentioned. [87] reported results obtained with GMM models using an EM learning procedure that improved only moderately over a raw acoustic features baseline. Chen et al. also reported results on language-mismatched posteriorgrams improving largely over the raw acoustic features baseline and close to a supervised baseline (17.2%, 17.5% and 16.3% ABX error rate for posterior grams trained respectively on Czech, Hungarian and Russian) in stark

[12] 12.0% ABX error rate for the Siamese Network vs. 12.1% for the supervised baseline in the within-speaker task on the Buckeye corpus.

[13] 28.6% AP on Switchboard for the GMM-based method from [87] vs. 46.9% AP for the Correspondence Autoencoder method from [102] and 54.9% AP for a modified version of the Siamese Networks learning whole-words embeddings rather than frame-level embeddings [91].

contrast with the earlier results from [34, 86] that we already reviewed and our own results from Chapter 4. Recently, an independent reproduction of Chen et al. DPGMM results has been attempted [107]. The results on the Buckeye corpus were not replicated, although the performance obtained remained the second best in the challenge (19.5% ABX error rate on the across-speaker task), while the results on the NCHLT corpus were actually improved, setting new records for the challenge both within and across speakers (but remaining below a supervised baseline). However, by training LDA features based on labels obtained from a first frame-level DPGMM clustering (thus remaining completely unsupervised) and performing a second round of DPGMM clustering on the basis of these features, the authors were able to beat the original results of Chen et al. on both corpora (even reaching the supervised baseline of 16.0% ABX error rate in the across speaker task on the Buckeye corpus). This idea of alternatively updating components to learn jointly and without supervision an acoustic model and acoustic features was also used in another recent study [108]. In this study the acoustic model used was not a pure GMM but rather used multiple HMM-GMM models and the acoustic features were not LDA features but bottleneck NN features. The results obtained (21.9% ABX error rate in the across-speaker task on the Buckeye corpus) were largely above the raw acoustic features baseline, but remained inferior to the results with DPGMM from [106, 107] and those obtained with NN trained based on top-down information from STD [101–103].

In summary, discrimination measures, and ABX-discriminability measures in particular, have been already largely used to evaluate speech processing systems operating without explicit supervision. They showed how the performance of such systems can be affected by the choice of acoustic features [35, 88, 96, 103], of an acoustic modeling method [87, 90, 97, 102, 105, 106] or the combination of both [107, 108]. An interesting finding was that excepted when using weak top-down constraints from STD techniques, NN-based models were consistently outperformed by GMM-based models. Thus it appears that, as of now, the impressive improvements brought about by the replacement of HMM-GMM system by NN-based systems in supervised speech processing do not translate to the unsupervised domain. Some of the proposed systems have yet to be compared to each other. In particular ISA features [88] and the model proposed by Lee et al. [90], while they yielded promising results in the AP metric on the TIMIT corpus, have not yet been compared with more recent proposals. Also, many of the proposed ideas are

complementary and there are many possible combinations that have not yet been tested. For example, using top-down cues from STD techniques could be combined with DPGMM modeling at the frame-level. However, the most pressing matter in the field is probably to investigate the source of the impressive performance of the DPGMM-based solutions. Indeed, this performance is in surprising contrast with earlier result using GMM-based acoustic models and appears hard to exactly replicate. It might be due for example to differences between EM-based and Dirichlet Process-based GMM optimization, the use of full-covariance or diagonal-covariance Gaussian components or the use of speaker normalization techniques such as VTLN on the features. Identifying the source of the discrepancy will shed light on important design principles for speech processing systems learning without supervision.

2.3 Application (ii): Modeling human or animal behavior in discrimination tasks

In this section, we consider a family of application of ABX-discriminability measures, where the objective is to model human or animal behavior in discrimination tasks. We discuss first the case where the task to be modeled is an ABX discrimination task in Section 2.3.1. Then, in Section 2.3.2, we consider the case of other types of discrimination tasks.

2.3.1 Modeling human or animal behavior in ABX discrimination tasks

In a typical ABX discrimination experiment, human or animal participants are presented with three stimuli A, B and X and their task is to indicate whether they perceive X as closer to A or to B. In this section, we propose a simple model for ABX discrimination experiments.

A simple model for the task follows directly from the way we defined ABX discriminability measures formally. In this model, participants first form internal representations a, b and x of the stimuli A, B and X to which they are exposed, then evaluate the similarity between A and X and between B and X based on some measure of dissimilarity d defined on the space of internal representations and finally answer A if $d(a, x) < d(b, x)$ and B otherwise.

A limit of this model is that it does not account explicitly for intrinsic sources of variability in the participants' responses. Some sources of variability can be accounted for implicitly without needing to alter the formalism. Indeed, the internal representations and the dissimilarity function

are supposed to be provided by a *perception* module (see Introduction chapter), which can incorporate some intrinsic variability mechanisms. For example, plausible limits on the resolution of the representations of the stimuli can be enforced by adding some artificial noise to them.

Intrinsic variability at the level of the decision process needs to be accounted for explicitly however. We propose to do this by considering that the participants answer A with a probability drawn from a Bernoulli distribution $\mathbf{b}(p)$, whose expectation is determined by the size of the difference between $d(b,x)$ and $d(a,x)$:

$$p = s(d(b,x) - d(a,x)),$$

where s is some sigmoid function. Parameters controlling the bias and variance of the response can also be introduced, for example by taking:

$$p = s\left(\frac{d(b,x) - d(a,x)}{\sigma} + \beta\right). \tag{2.1}$$

The simpler model, without randomness at the level of the decision process, is retrieved for a model without bias ($\beta = 0$) by taking $s = \mathbb{1}$, where $\mathbb{1}$ is the indicator function, which can be seen as the limit of other sigmoid functions when the variance parameter σ tends toward 0 ($\mathbb{1}(x)$ is equal to 0 if $x < 0$, 0.5 if $x = 0$ and 1 if $x > 0$).

What motivation is there for proposing this model? We will show that it can be seen as a generalization of a model described by McMillan & Creelman ([109], p.233). The interest of the generalization is that it provides models that are more appropriate for tasks involving complex stimuli that cannot be seen as being distributed along a single dimension (such as speech signals).

McMillan & Creelman propose a *Roving ABX* model for the ABX match-to-sample task within the framework of Signal Detection Theory. In this model, the participants are supposed to represent each stimulus as a point on some internal psychological dimension. Let us note $\hat{a}$, $\hat{b}$ and $\hat{x}$ the internal representations of stimuli A, B and X. $\hat{a}$ and $\hat{x}$ are supposed to be independent samples from a Gaussian distribution $\mathcal{N}(a, \sigma^2)$ and $\hat{b}$ is supposed to be a sample from a Gaussian distribution $\mathcal{N}(b, \sigma^2)$ (mutually independent from $\hat{a}$ and $\hat{x}$). The model predicts that an unbiased observer will answer A if $|\hat{a} - \hat{x}| < |\hat{b} - \hat{x}|$ and B otherwise. In this model, the

expected percentage of A answers (i.e. correct answers in this case) for the unbiased observer is:

$$p(d') = \Phi\left(\frac{d'}{\sqrt{6}}\right)\Phi\left(\frac{d'}{\sqrt{2}}\right) + \Phi\left(-\frac{d'}{\sqrt{6}}\right)\Phi\left(-\frac{d'}{\sqrt{2}}\right),$$

where $d' = \frac{|b-a|}{\sigma}$ and Φ is the standard normal cumulative distribution function.

To see the link with our model, we need to make two modifications. First, we consider that $\hat{x}$ is taken from a distribution $\mathcal{N}(x, \sigma^2)$ where x can be different from a and b. Second, we consider a slightly different task: an AXBX task, where A and X are presented first and then B and X again. The participant is asked to answer whether X was closer to A or B as in the classical ABX match-to-sample task. The goal of this second change is to avoid asymmetries between the expected performance for fixed values of $|a-x|$ and $|b-x|$ depending on whether a and b are on the same side of x or on opposite sides. These asymmetries are materialized by the presence of two different factors, $\sqrt{6}$ and $\sqrt{2}$ in the expression of $p(d')$ above and stem from the correlations that are introduced by using the same sample $\hat{x}$ from $\mathcal{N}(x, \sigma^2)$ to compute $\hat{a} - \hat{x}$ and $\hat{b} - \hat{x}$.

In an AXBX task, we assume that we obtain two independent draws $\hat{x}_1$ and $\hat{x}_2$ from $\mathcal{N}(x, \sigma^2)$ corresponding to the two different presentations of X. Then using the decision rule: answer A if $|\hat{a} - \hat{x}_1| < |\hat{b} - \hat{x}_2|$, else B, the expected percentage of A answers for the unbiased observer can now be shown to be:

$$p(S, \Delta) = \Phi\left(\frac{\Delta}{2}\right)\Phi\left(\frac{S}{2}\right) + \Phi\left(-\frac{\Delta}{2}\right)\Phi\left(-\frac{S}{2}\right),$$

where $S = \frac{|b-x|+|a-x|}{\sigma}$ and $\Delta = \frac{|b-x|-|a-x|}{\sigma}$.

By looking at a plot of $p(S, \Delta)$ (see Figure 2.1), we see that p does not depend a lot on S and is well approximated as a sigmoid function of Δ, i.e. there exists a sigmoid function s such that $p(S, \Delta) \approx s(\Delta)$. Then our more general model, which specifies that $p = s(d(b, x) - d(a, x))$, can be obtained by replacing $\frac{|b-x|}{\sigma}$ and $\frac{|a-x|}{\sigma}$ in the expression of Δ by the much more general quantities $d(b, x)$ and $d(a, x)$. In this way, the constraint that a, b and x are real numbers is lifted and all we need is to be able to compute dissimilarities between them. This allows using complex stimuli for which a comparison on a single dimension would not make much sense.

Let us note that both McMillan & Creelman's model and ours remain quite abstract models of the task and that there are some well-documented effects for which they do not account

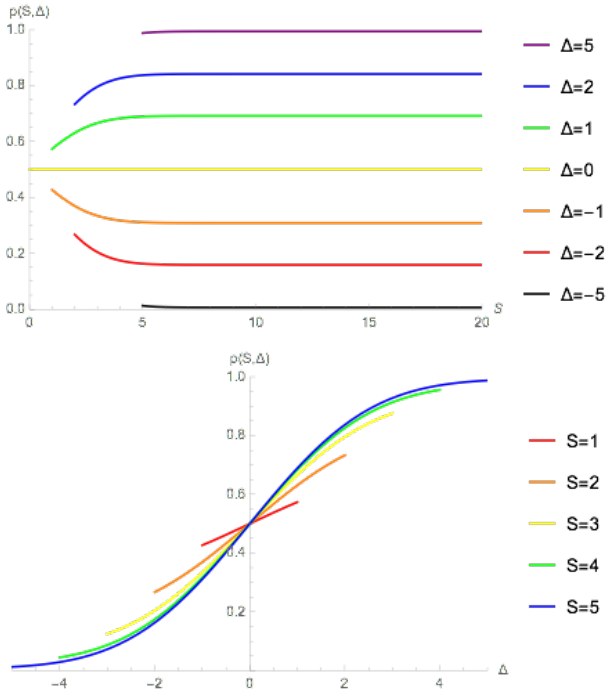


Figure 2.1: Plot of $p(S, \Delta)$.The plot bounds are determined by the facts that $-S \leq \Delta \leq S$ and $S \geq 0$. The top plot, representing $p(S, \Delta)$ as a function of S for several values of Δ, shows that p does not vary a lot as a function of S. The bottom plot, representing $p(S, \Delta)$ as a function of Δ for several values of S, shows that p is well approximated as a sigmoid function of Δ.

explicitly. For example, for sequentially presented stimuli, the inter-stimuli interval is known to affect the performance of participants, as is the total number of different stimuli used in the experiment (see for example [109], chapter 9).

In conclusion, our extension of the McMillan model enables us to derive predictions for human or animal performance in an AXBX task from a *perception* model (i.e. a mapping from stimuli to representations together with a notion of dissimilarity on the space of representations). Of course, these predictions would have to be tested empirically (just as the predictions of Signal Detection Theory can and have been tested empirically). In practice, we can use formula 2.1 to estimate an appropriate shape for the sigmoidal function and its bias and variance parameters from experimental measurements. The fitted values can then be used to compute the agreement of different models of perception with the empirical observations.

2.3.2 Modeling human or animal behavior in other discrimination tasks

ABX discrimination tasks with humans and animals are closely related to other discrimination tasks, in particular to those from the family of *classification designs for discrimination* ([109] chapter 9), which also includes the same-different and oddity experimental tasks. Signal Detection Theory models for these different tasks predict systematic relationships between the expected results, which have been confirmed empirically, at least partially, in numerous experiments, using a wide variety of experimental stimuli ([109], chapter 10). These results were obtained with stimuli that could be assumed to be distributed along a single dimension, but generalizations of the Signal Detection Theory models that are suitable for more complicated ensembles of stimuli can be developed for the different tasks, following the same approach as in the previous section. This suggests that experimental results in same-different or oddity discrimination tasks should agree qualitatively with the results in an ABX task even with more complex stimuli.

Why is this interesting? The link with same-different and oddity experimental paradigms is interesting because many perception experiments with infants are based on these paradigms. This means that, even though infants are not able to perform ABX tasks, ABX discriminability measures might be useful to predict their behavior. This is in contrast with classification or clustering experimental paradigms, which have no direct relationship with same-different or

oddity paradigms. Of course, when it is possible to use directly a model of a same-different or oddity task, there is no reason to use ABX discriminability measures. However, existing models for these tasks typically require setting a threshold parameter, which corresponds to a point beyond which two stimuli are not considered the same anymore or to the point beyond which a stimulus is considered an oddball. By contrast, the most simple model that we proposed for ABX tasks does not rely on any parameter (given a perception model). Thus if suitable data for fitting the parameters of more direct models is not available, ABX discriminability measures offer a potential solution for predicting qualitatively the behavior of infants in discrimination tasks.

This can be useful as a heuristic for planning informative experiments. For example, if a set of interesting perception models for phonetic category acquisition has been identified, phonetic contrasts for which the ABX discriminability predicted by the different models are the most different can be singled out. An empirical test to decide between the different models can then be designed on the basis of these maximally informative contrasts. More generally if the raw experimental data is not available or does not allow the fitting of a threshold parameter, ABX discriminability measures might be useful.

2.4 Application (iii): Providing descriptive measurements of datasets with category labels

In the third and last type of application we consider, one or several datasets annotated with category labels for some category structures of interest are supposed to be available and the goal is to get a qualitative understanding of how the category structures are represented in the datasets. The datasets and category structures can be anything. For example, it could be a collection of images of human faces annotated with category labels for the identity of the person, their gender, their age, personality traits, the emotion they display, etc. Or it could be videos of behaving agents with annotations for the type of action in which the agents are engaging, the identity of the agents, the type of location or the objects present in the scene. In the remainder of this section, we illustrate the interest of ABX-discriminability measures in the specific case where the datasets considered are transcribed corpora of speech recordings.

Phonetic properties of speech, are often studied by extracting phonetic measurements, such as

formants or VOT values, from portions of interest of speech recordings. These measurements are typically extracted by hand by trained phoneticians, relying on their subjective judgments. This is very costly in time and prevents available large corpora of recorded speech to be exploited to their full potential. In this section, we present an alternative approach, based on ABX discriminability measures, that do not suffer from this problem.

Traditional phonetic measurements have to be derived by hand because there is a lack of reliable methods for extracting them automatically in a way that can handle the variability present in natural speech. One approach to solving this problem is to design better methods for automatic extraction of phonetic measurements, and there has been some work in this direction (see for example [110, 111]). We propose a different approach: replacing traditional phonetic measurements by featural representations of speech as used in Automatic Speech Recognition (ASR) systems and speech technology in general (e.g. MFC or PLP coefficients [18, 112] or various kinds of spectrograms [113, 114]). The main benefit of this approach is that these featural representations are designed to be reliably extractable from speech signal in a fully automatic fashion. The main difficulty is that these high-dimensional representations are much harder to interpret and manipulate than traditional phonetic measurements. To resolve this difficulty and obtain measurements that are both reliably extractable in an automatic fashion and easily interpretable, we propose to perform ABX discriminability measures on the basis of the featural representations.

To illustrate the approach, we present two examples. A systematic comparison of the discriminability of phonetic contrasts in Infant-Directed Speech (IDS) and Adult-Directed Speech (ADS) in Japanese in Section 2.4.1. And a systematic study of the effect of coarticulation on the discriminability of phonetic segments in Japanese and English in Section 2.4.2. A brief conclusion is provided in Section 2.4.3.

2.4.1 Are phones more distinct in Infant-Directed Speech?

The experiments and results summarized in this section were performed in collaboration with A. Martin, M. Versteegh, K. Miyazawa, R. Mazuka and A. Cristià and published in [115]. R. Mazuka oversaw the collection and coding of the corpus. K. Miyazawa provided key analyses. A. Martin wrote the syllabification algorithms. M. Versteegh wrote the feature-extraction algorithms. T. Schatz designed the ABX tasks. E. Dupoux carried out the acoustical analyses. A Cristià

carried out the statistical analyses. A. Martin and A. Cristià produced the first draft and all authors contributed to the definition of the research question, the methodological approach and the writing of the manuscript.

The *hyperarticulation of IDS* hypothesis for phonetic categories states that in order to facilitate phonetic category acquisition by infants, caregivers produce speech with more distinct phonetic contrasts when they talk to infants than when they talk to adults [116]. The RIKEN Japanese Mother-Infant Conversation Corpus [117, 118] is particularly appropriate for testing this hypothesis: it contains spontaneous speech from 22 mothers in two conditions: showing picture books and playing with their child and speaking with an adult experimenter.

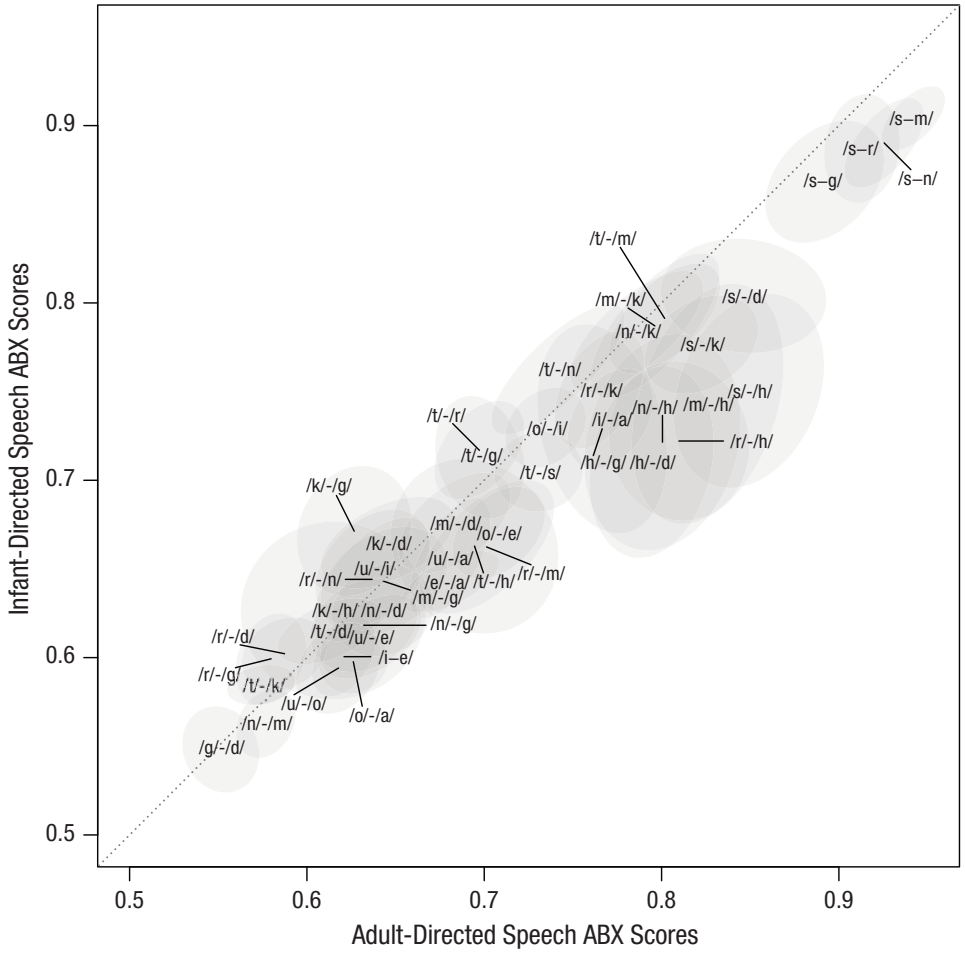


Figure 2.2: Average ABX discriminability for IDS as a function of ABX discriminability for ADS for each contrast between syllable onsets for sentence-medial syllables (collapsed across contexts). For each contrast, the average across speakers is either at the center of the contrast label or at the endpoint of the line leading from the contrast label. The eclipsed indicate 95% confidence intervals over the 22 speakers. The dotted gray line indicates where points would fall if the averages where equal. /u/ stands for /ɯ/. Figure reproduced from [115].

Operationalizing the hypothesis in terms of an ABX discrimination task is straightforward: if we compute ABX discriminability scores for each possible phonetic contrast from the IDS and ADS speech samples separately, then, for the *hyperarticulation* hypothesis to hold, the difference of the matched scores (IDS−ADS) should be on average positive. Computing scores for minimal-pair instead of words is important to avoid confusing effects of the discriminability of phones with effects of the different lexicon used in IDS and ADS (for example one could imagine that even without any difference in phonetic discriminability between IDS and ADS, frequent words in IDS involve easier contrasts more often than frequent words in ADS). To compute a score for each phonetic contrast, the approach we followed was to segment the speech signal into syllables and compute an ABX discriminability measure ON syllable onset BY the rest of the syllable (the context), the position of the sentence in the utterance (initial, medial, final or isolated), the register (IDS or ADS) and the talker. For example, an ABX triplet in the task could be:

A	B	X
$/\text{saN}/_i^{A,T_1}$	$/\text{taN}/_i^{A,T_1}$	$/\text{saN}/_i^{A,T_1}$

where $_i$ indicates that the syllable occurs sentence-initially, ^{A} indicates that the register is ADS and ^{T_1} indicates that the syllables was uttered by talker T_1. In this example, the measure is ON the /s/-/t/ phonetic contrast, BY the /aN/ context, the initial sentence-position and the ADS register.

ABX discriminability measures were computed on the basis of sequences of spectral frames of duration 25ms computed every 10ms, with a Mel warping of the frequency scale and a cubic root compression of the dynamic range. The reason for choosing this representation format was, as we will see in chapter 3, that it can be interpreted as a crude model of speech perception by the baby at birth. The idea being that we are interested in the characteristic of IDS from the point of view of the child. Using MFC coefficients instead did not change the pattern of results observed. As with other experiments with spectral or cepstral-based representations dissimilarities were computed through Dynamic Time Warping based on a frame-to-frame cosine distance.

Results for phonetic contrasts for which enough data was available in the corpus for reliably estimating their ABX discriminability are shown in Figure 2.2. The results are shown for sentence-medial syllables and show that most phonetic contrasts are more discriminable in ADS than in IDS. This tendency was found to be significant through a permutation test of the

IDS and ADS conditions within each talker. Similar results were obtained for sentence-initial syllables and no significant difference was found for sentence-final syllables. Thus, our results do not support the *hyperarticulation* hypothesis. To the contrary, they indicate that phonetic contrasts are slightly less discriminable on average in IDS than in ADS. These results are in line with recent empirical evidence and theoretical analyses proposing alternative accounts for the specific phonetic and prosodic properties of IDS, highlighting its *communicative* function [119–122].

In the end, ABX discriminability measures allowed us to perform a comprehensive test of the *hyperarticulation of IDS* hypothesis, with an unprecedented coverage of 10 vowel and 36 consonant contrasts. This was achieved by exploiting a corpus containing a total of 14 hours of speech (11 hours of IDS and 3 hours of ADS), using manually annotated phone-level transcriptions and time-alignments as well as sentence boundaries.

2.4.2 Measuring the impact of coarticulation on phone discriminability

The acoustic realizations of a given phonetic segment are typically affected by the preceding and following phonetic segments, a phenomenon called coarticulation (see for instance [123] pp. 70-71). For example, the acoustic realizations of the American English vowel /ɪ/ might be systematically different according to whether it occurs in the context of a preceding /m/ and a following /s/ as in the word *miss* or in the context of a preceding /t/ and a following /p/ as in the word *tip*. In this section, we show how ABX discriminability measures can be used to study these coarticulation effects quantitatively and in a systematic fashion. We begin by defining a measure of interest for studying coarticulation effects in Section 2.4.2.1, that we then apply to two large corpora of recorded speech, one in American English and one in Japanese, in Section 2.4.2.2.

2.4.2.1 Definition of the measure

Our idea is to consider a phonetic contrast occurring in a given language and to measure how well it can be discriminated, on the one hand, based on acoustic realizations occurring in the same phonetic context and, on the other hand, based on acoustic realizations occurring in different phonetic contexts. If the two phonetic segments involved are completely unaffected

by coarticulation, then we expect to see no difference between the scores in the two cases, but in the presence of coarticulation effects, the two score can be different. We thus propose to take the difference between these scores as a measure of the impact of coarticulation on the discriminability between phonetic segments. More precisely, we consider three different ABX tasks. First, the *within context* task (WT), performed ON phonetic segment, BY talker, preceding context and following context. An ABX triplet in this task could be for example:

A	B	X
$/\mathrm{i}/^{T_1}_{b\_t}$	$/\mathrm{u}/^{T_1}_{b\_t}$	$/\mathrm{i}/^{T_1}_{b\_t}$

where ${}_{b\_t}$ indicates a segment preceded by a /b/ and followed by /t/ and ^{T_1} indicates a segment pronounced by speaker T_1. Second, the *across preceding context* task (PT), performed ON phonetic segment, ACROSS preceding context BY talker and following context. An ABX triplet in this task could be for example:

A	B	X
$/\mathrm{i}/^{T_1}_{b\_t}$	$/\mathrm{u}/^{T_1}_{b\_t}$	$/\mathrm{i}/^{T_1}_{s\_t}$

Third, the *across following context* task (FT), performed ON phonetic segment, ACROSS following context BY talker and preceding context. An ABX triplet in this task could be for example:

A	B	X
$/\mathrm{i}/^{T_1}_{b\_t}$	$/\mathrm{u}/^{T_1}_{b\_t}$	$/\mathrm{i}/^{T_1}_{b\_n}$

For each task and each phonetic contrast, we compute a summary ABX score as follows. We start from ABX discriminability measures for each combination of talker, preceding context(s), following context(s) and phonetic contrast. First, we average out the talkers to obtain score for each combination of preceding context(s), following context(s) and phonetic contrast. Second, we average out the phonetic contexts to obtain a score for each phonetic contrast. Let us note $s^{WT}(p_1, p_2)$, $s^{PT}(p_1, p_2)$ and $s^{FT}(p_1, p_2)$ the ABX scores obtained in this fashion for the p_1/p_2 phonetic contrast in the WT, PT and FT task respectively. Our main measure for each phonetic contrast is then the *coarticulation* score:

$$s_c(p_1, p_2) = s^{WT}(p_1, p_2) - \frac{s^{PT}(p_1, p_2) + s^{FT}(p_1, p_2)}{2}.$$

Finally, we compute a score for each vowel by averaging the scores for each vocalic phonetic contrast involving that vowel and for each consonant by averaging the scores for each consonantal contrast involving that consonant.

2.4.2.2 Application

We present results obtained by computing the measures defined above on speech stimuli from the *Wall Street Journal* corpus [20] and from the *Corpus of Spontaneous Japanese* [124]. We used a subset of the *Wall Street Journal* corpus [20] containing recordings from 20 native American English speakers reading news articles from the Wall Street Journal and containing a total of 242.654 phonetic segments for a duration of approximately 6 hours. For an experiment focusing on the /s/-/ʃ/ contrast in the context of the /u/ and /i/ vowels (see below), we also used a larger subset of the corpus containing speech from 169 speakers. The corpus was designed to facilitate the training of large vocabulary speech recognition systems and the recordings have been checked by the corpus providers for hesitations and pronunciation errors to ensure a good match between the text of the article and the recordings. Phonetic transcriptions were obtained using the CMU phonetic dictionary of American English (`http://www.speech.cs.cmu.edu/cgi-bin/cmudict`) and a phone-level forced-alignment was obtained using a speaker-adapted triphone HMM-GMM speech recognizer trained on the corpus. We used a subset from the *Corpus of Spontaneous Japanese* containing audio recordings from 39 native speakers of Japanese speaking spontaneously about an episode of their life in front of a small audience. This subset contains a total of 277.832 phonetic segments for a duration of approximately 6 hours. Manually-checked phonetic transcriptions were provided with the recordings for the considered subset and a phone-level forced-alignment was obtained using a speaker-adapted triphone HMM-GMM speech recognizer trained on the corpus.

The audio recordings for both corpora were coded as a sequence of MFC coefficients vectors taken every 10ms. Each phonetic segment was represented as the sequence of MFC coefficients vectors that occurred between the beginning and the end of that segment as specified by the phone-level time-alignments. DTW on a frame-level cosine distance was used as the distance function. We ignored word-boundaries and syllable structure in the formation of ABX triplets and the context of phones at the beginning and end of a sentence was marked using a special

silence symbol *sil*. For example, the sentence *Some tea* is considered as containing the following phone/context pairs: /s, sil_ʌ/, /ʌ, s_m/, /m, ʌ_t/, /t, m_i/ and /i, t_sil/.

Segment	Coarticulation score (%)
iː	5.6
äː	6.5
oː	7.0
ä	7.3
eː	8.1
o	8.1
e	8.6
i	9.4
ɯː	10.3
ɯ	11.5

Table 2.1: Coarticulation scores for Japanese vowels, sorted in increasing order.

Segment	Coarticulation score (%)
iː	5.3
eɪ	7.0
ɑː	7.4
ɔː	7.5
aʊ	7.8
ɝ	8.0
oʊ	9.1
aɪ	9.3
ɔɪ	9.4
æ	9.5
ɪ	9.6
ɛ	9.9
ʌ	10.0
uː	10.3
ʊ	13.5

Table 2.2: Coarticulation scores for American English vowels, sorted in increasing order.

The coarticulation scores obtained for each phonetic segment of each language are reported in Tables 2.1, 2.2, 2.3 and 2.4. Looking first at the pattern of results for vowels, we see that in Japanese the long version of each vowel has a lower coarticulation score than the short version. A similar pattern is also observed for matching pairs of tense and lax vowels in American English: /uː/ has a lower score than /ʊ/, /eɪ/ has a lower score than /ɛ/ and /iː/ has a lower score than i. The distribution of scores for short and long vowels in Japanese and tense and lax vowels in American English are shown in the *Vowels* panel of Figure 2.3, where it appears clearly that the discriminability of short/lax vowels suffers more from coarticulation than the discriminability of long/tense vowels. This could be explained, for example, if the periods of transition with the previous and the following segments, where coarticulation effects are expected to be strongest, do not become longer with longer segments, so that a larger part of the longer segment remains immune to coarticulation. Another interesting result is that for both languages the two vowels

Segment	Coarticulation score (%)
ɸ	4.2
ɕː	4.4
s	4.5
ɕ	5.9
z	6.0
ʑ	6.4
sː	6.6
j	7.8
tː	7.9
t	7.9
b	8.1
d	8.4
p	8.8
m	8.9
r	8.9
ɴ	9.2
pː	9.2
w	9.8
g	9.9
n	10.0
k	11.7
h	12.6
ʔ	13.7
kː	14.3

Table 2.3: Coarticulation scores for Japanese consonants, sorted in increasing order.

Segment	Coarticulation score (%)
ʃ	1.6
s	1.6
z	2.4
ʒ	2.7
tʃ	3.1
f	4.5
dʒ	4.8
j	4.9
w	6.4
r	6.4
b	6.9
l	7.1
ŋ	7.1
v	7.5
d	7.7
m	7.7
g	7.8
p	8.2
n	8.3
θ	8.4
ð	8.9
h	9.8
t	9.9
k	10.7

Table 2.4: Coarticulation scores for American English consonants, sorted in increasing order.

that appear to suffer the most from coarticulation are the long/tense and short/lax close back vowels (/uː/ and /ʊ/ for English, and /ɯː/ and /ɯ/ for Japanese) and the vowel that appears to suffer the least is the close front vowel /iː/.

For consonants, the most salient pattern we observe in the result is that most fricatives have very low coarticulation scores, while most stops have rather high scores. We grouped the coarticulation scores according to the manner of articulation of the different segments and represented the distribution of scores for the different groups in the *Consonants* panel of Figure 2.3. This analysis shows that on average fricatives have lower coarticulation scores than other

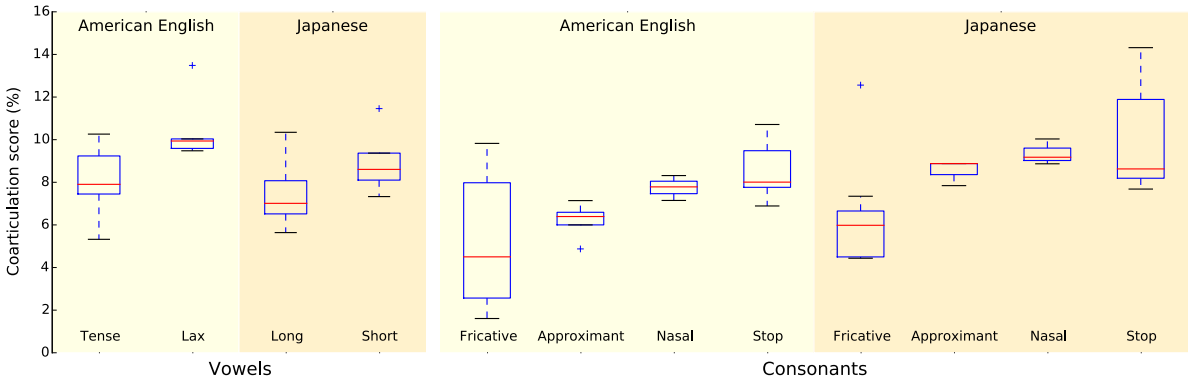


Figure 2.3: Boxplot of the distributions of the coarticulation scores obtained for different class of vowels and consonants in American English and in Japanese. For consonants, laterals were pooled with approximants and and affricates with fricatives.

groups of segments. Approximants appear to have slightly lower scores on average than nasals and stops although the difference is less marked, especially in Japanese. Nasals and stops have roughly the same average coarticulation scores. Also, all nasals and all approximant appear to have similar coarticulation scores, while there is much more variability in the scores of the different fricatives and stops. Looking more closely at fricatives, it appears that the distribution of coarticulation scores is bimodal. In both languages all the fricatives have coarticulation scores that are lower than the lowest score for a non-fricative segment, except for /h/ in Japanese and /h/, /θ/ and /ð/ in English, which have coarticulation scores among the 5 highest for consonants in their respective languages. For English stops, voiced stops have globally lower coarticulation scores than voiceless stops and, for a fixed value of voicing, stops with a more anterior place of articulation have lower coarticulation scores. For Japanese stops, the pattern is different. In particular, the /t/ (and geminate /tː/) segment has the lowest coarticulation score of all stops, whereas it had the second highest score in English. For the rest of the stops, voicing and anteriority of the place of articulation are still associated with lower coarticulation scores, but it is not the case anymore that the highest score for voiced stops is lower than the lowest score for voiceless stops. In both languages, the consonant with the highest score is voiceless velar stop (a geminate one for Japanese).

The very low coarticulation scores for fricatives that we observe might seem surprising given that coarticulation effects have been reported for fricatives. For example, the so-called *sushi* effect [125], suggests that the frontier between the /s/ and /ʃ/ categories in American English

is shifted in the direction of /ʃ/ when the following segment is a /uː/ and in the direction of /s/ when the following segment is a /iː/. We now ask whether this effect is compatible with our measurements. The *sushi* effect predicts that /s/ and /ʃ/ segments should be closer on average when the /s/ segment is followed by an /uː/ and the /ʃ/ segment is followed by an /iː/ than when the /s/ segment is followed by an /iː/ and the /ʃ/ is followed by an /uː/. We can formulate this in terms of (non-symetrized) ABX discriminability measures, using the notations from Chapter 1 as:

$$\mathcal{D}_{abx}\left(\mathbb{P}(\text{s, i}), \mathbb{P}(\text{ʃ, i}), \mathbb{P}(\text{s, u})\right) < \mathcal{D}_{abx}\left(\mathbb{P}(\text{s, u}), \mathbb{P}(\text{ʃ, u}), \mathbb{P}(\text{s, i})\right),$$

and symmetrically:

$$\mathcal{D}_{abx}\left(\mathbb{P}(\text{ʃ, u}), \mathbb{P}(\text{s, u}), \mathbb{P}(\text{ʃ, i})\right) < \mathcal{D}_{abx}\left(\mathbb{P}(\text{ʃ, i}), \mathbb{P}(\text{s, i}), \mathbb{P}(\text{ʃ, u})\right),$$

where $\mathbb{P}(s_1, s_2)$ stands for the probability distribution of the representation (i.e. here a sequence of MFC coefficient vectors) of segment s_1 when it is followed by segment s_2. To obtain enough exemplars to reliably estimate these ABX scores, we used a larger subset of the Wall Street Journal corpus, containing 169 speakers. This yielded a total of 7998 /s/ and S/ phonetic segments occurring before /iː/ or /uː/. The estimated ABX scores are compatible with the predictions (still using the notations from Chapter 1):

$$\hat{\theta}\left(\mathbb{P}(\text{s, i}), \mathbb{P}(\text{ʃ, i}), \mathbb{P}(\text{s, u})\right) = 96.8 < \hat{\theta}\left(\mathbb{P}(\text{s, u}), \mathbb{P}(\text{ʃ, u}), \mathbb{P}(\text{s, i})\right) = 98.1,$$

and:

$$\hat{\theta}\left(\mathbb{P}(\text{ʃ, u}), \mathbb{P}(\text{s, u}), \mathbb{P}(\text{ʃ, i})\right) = 96.5 < \hat{\theta}\left(\mathbb{P}(\text{ʃ, i}), \mathbb{P}(\text{s, i}), \mathbb{P}(\text{ʃ, u})\right) = 100.0.$$

So, the weak impact of coarticulation on the discriminability of fricatives when compared to other kinds of segments, does not appear incompatible with the existence of coarticulation effects on fricatives. Since we are measuring the impact of coarticulation on discriminability, the absolute size of the coarticulation effects is not as important as their relative size with respect to the separation between phonetic categories. Indeed, if two categories are very well separated, they can remain highly discriminable even in the presence of large coarticulation effects, while

categories that are already hard to discriminate can suffer even from small coarticulation effects. This suggests that our coarticulation scores should be correlated with ABX discriminability measures on the WT task. This is indeed the case for both English ($r = 0.73$) and Japanese ($r = 0.55$). An interesting follow-up would be to try to regress this correlation in order to decompose our coarticulation scores for each phonetic contrast into two components: a component depending only on the absolute discriminability between the categories and a residual component specific to the contrast under study. We just scratched the surface of the possibility offered by ABX discriminability measures and there are many ways in which the study we presented could be extended. First, we provided a rather descriptive analysis of the results and did not attempt, given time constraints and our limited expertise on the topic, to systematically relate our results to the existing literature on coarticulation effects. Hopefully, this is still sufficient to demonstrate the interest of the method. It would also be interesting to study potential undesirable effects related to the use of phonetic transcriptions based on phonetic dictionaries and/or the use of forced alignments obtained from automatic speech recognizers. This could be readily done using corpora like the Corpus of Spontaneous Japanese, for which manual transcription and alignments are available. Other possible extensions include looking at other specific coarticulation effects beyond the *sushi* effect, studying separately regressive and progressive coarticulation effects (corresponding to the FT and PT tasks respectively), looking at gender effects, taking into account word-boundaries or syllable structure or lexical stress, studying other languages...

2.4.3 Conclusion on phonetic measurements

In the two previous sections, we showed how, given a perception model (i.e. a representation and a similarity function), ABX discriminability measures can be used to obtain phonetic measurements on large corpora of recorded speech, that are much more systematic than traditional hand-made-measurements-based alternatives in at least two respects. First, the automaticity of the procedure allows to exploit larger corpora, cover more contrasts, study more sources of variability... Second, representations that cover the acoustic space more systematically than traditional phonetic measurements can be used. We obtained interesting and interpretable results with this method in situations involving the comparison of several corpora in the same task and in situations involving the comparison of the results of several tasks applied to the same corpus. We were able to describe global tendencies and to retrieve particular effects.

Part II

Applications to models of early phonetic category acquisition

Chapter 3

Modeling phonetic category perception at birth

Contents

In this chapter, we introduce a framework for designing and evaluating computational models of phonetic category perception at birth. We then consider classical methods for speech features extraction in Automatic Speech Recognition (ASR) as potential models of phonetic category perception at birth within this framework. The main motivation behind this work is to obtain models of phonetic category perception at birth that can be used as a natural starting point for

modeling the subsequent process of language-specific phonetic category *acquisition* [1].

The method we propose is rather indirect because of the inherent difficulties associated with the study of the behavior or cerebral activity of neonates, which severely limit the amount of direct empirical evidence available to suggest or evaluate potential computational models. We follow a two-step approach. In the first step, we exploit knowledge about the human auditory system and the specific nature of the speech signal to suggest candidate models. In the second step, a variety of ABX-discriminability measures is computed to evaluate how effective the candidate models are in representing phonetic categories *from the point of view of a system learning without explicit supervision*.

The rationale for our two-step approach is detailed in Section 3.1. We then proceed to a practical application. In Section 3.2 (step 1), we introduce and motivate a family of candidate models suggested by ASR practice, which includes classical MFC [18] and a version of PLP [112, 126] coefficients as special cases. In Section 3.3 (step 2), we evaluate and compare how well the different models in this family can discriminate phonetic categories. We discuss the results and conclude in Section 3.4.

3.1 A two-step approach

We follow a two-step approach, where the first step serves to suggest potential computational models of phonetic category perception at birth, which are then evaluated in the second step. Direct evidence regarding the perception of phonetic categories at birth is scarce. We know that neonates can perceive differences between most of the speech sounds that are contrastive in the languages of the world [2, 127–139]. We also know that newborns show evidence of categorical perception in the case of certain consonant contrasts with category boundaries matching those commonly observed in the languages of the world [127–129]. Unfortunately, these results do not appear sufficient in themselves to suggest specific computational models for phonetic category perception at birth (step 1). Let us discuss why in the case of the results about the discriminability of contrastive speech sounds first. These results are essentially binary: evidence that a particular contrast can be discriminated by newborns is either found or not. Because measurements on neonates are both difficult to acquire and very noisy, more detailed results, such as graded measures of the relative difficulty in discriminating different contrasts are usually not

available. Since, in addition, stimuli with little or no variability are used in the experiments, we expect any model representing physically different sounds differently to be able to predict the positive results observed. Therefore, while the results about the discriminability of contrastive speech sounds can rule out models that would not represent the speech signal in sufficient details, they cannot be used to reject models that would represent the signal in too many details. For example, a computational model representing speech directly as an acoustic pressure waveform is likely to be sufficient to account for these results. This makes them essentially useless for both steps in our approach. The results about categorical perception of certain consonant contrasts appear more constraining and it would be interesting to use them to evaluate candidate models (step 2), but we do not see any obvious way to employ them to suggest specific models (step 1).

Thus, at least for the first step, we need to turn to less direct sources of inspiration. We consider two independent heuristics. According to the first heuristic, speech processing at birth, at least in the auditory modality, is mediated by generic auditory processes. The motivation for this heuristic comes from a series of results showing that the speech perception effects observed in neonates are also found in evolutionary close animals [140–143]. This includes the discriminability and categorical perception effects for phonetic categories that we mentioned above [144–148]. According to the second heuristic, speech processing at birth is already well-suited to decipher the linguistic content of the speech signal (although not in a language-specific way obviously). The rationale for this second heuristic is simply that the ability to decode the linguistic content of the speech signal is likely to be a significant evolutionary advantage, thus having been optimized through evolutionary processes.

These two heuristics might seem to contradict each other at first glance, because the former considers that speech processing in neonates is mediated by *generic* auditory processes, while the latter considers that it is adapted to the *specific* nature of the speech signal. Nevertheless, as we explain in the following, to the best of our knowledge both heuristics are compatible with the existing empirical evidence. This means that, whether they are contradictory or not, there appears to be no reason to prefer one over the other and both are worth exploring. Furthermore, as we will see, there is actually evidence that both heuristics are to some extent compatible and complementary. We motivated the second heuristic solely on logical grounds, without providing specific empirical evidence, but we did provide empirical evidence in favor of the first. There-

fore, we need to check that the empirical evidence we provided to back the first heuristic, about the similar perceptual abilities of newborns and animals, is compatible with the second heuristic, which states that newborn's perception is optimized for speech processing. The apparent paradox is that, by combining the second heuristic with the empirical evidence supporting the first heuristic, we can conclude that evolutionary close animals are well-endowed to decode the linguistic content of the speech signal, even though it is hard to imagine how efficient processing of human speech could be an evolutionary advantage to these animals. However, this could simply happen as a result of an evolution of the speech signal to adapt to the human auditory system, and thus, indirectly, to the very similar auditory systems of evolutionary close animals. There are at least two lines of empirical evidence supporting this hypothesis. First, in terms of evolutionary adaptations to speech in humans, there is clear evidence of changes at the level of the speech production apparatus [149, 150] but not at the level of the speech perception apparatus. Indeed, the general organization of the ear and the subcortical auditory pathways is similar across all mammals (see for example [151], Chapter 28) and there is to our knowledge no evidence for subcortical specialization for speech processing even in human adults (see for example [152], Chapter 4). This suggests that evolution of the human vocal tract and more generally of the whole speech production apparatus resulted in the ability to produce a signal, speech, that is shaped to match pre-existing auditory processing abilities. Second, in a famous computational study, Lewicki [153] obtained very similar results when optimizing a filterbank for sparse decomposition of either natural sounds or speech sounds. In both cases, the properties of the resulting filterbank accurately matched known properties of auditory processing at the level of the inner ear. This suggests that the human ear, as well as the one of evolutionary close animals, is optimized for processing natural sounds and that speech sounds evolved to match the properties of natural sounds. These two lines of evidence suggest, at the very least, that it is not absurd to consider that the speech signal evolved to match pre-existing perceptual abilities. From this, we can conclude two things. First, that none of the two heuristics we considered as inspiration for models of phonetic categories perception at birth can be rejected on the basis of the available empirical evidence. This leads us to pursue both heuristics in parallel. Second, that it is plausible that both heuristics, although they are completely independent, will provide compatible results. If this proves to be the case, even partially, it would mean that both heuris-

tics can be combined to obtain more reliable models (certain aspects of the models might be independently corroborated by the two heuristics) with a wider scope (certain aspects of the models might be specified by only one of the two heuristics). In our study of an ASR pipeline in Section 3.2, we indeed observe that both heuristics provide compatible and complementary results.

Now that we have introduced these two heuristics and how they relate to each other, let us explain how they can be exploited in our quest for potential models of phonetic category perception at birth. The main reason the first heuristic is particularly useful is that the auditory system is known to be mature at the time of birth in many respects [154]. This means that a large part of the vast literature on auditory physiology and psychology in adults is relevant to our problem. The main reason the second heuristic is particularly useful is that it suggests that the vast amount of knowledge accumulated in fields studying the nature of the speech signal, its production and its decoding is also potentially relevant. This includes in particular results from phonetic science, signal processing and speech engineering.

In summary, the first step of our method consists in proposing potential computational models of phonetic category perception at birth based on converging results from auditory physiology and psychology, phonetics, signal processing and speech engineering. Let us now describe and motivate the second step of our method. The first step typically yields a variety of potential models, each with many loose parameters. How do we decide between them? The most obvious approach would be to try and confront the different models to the available empirical data on phonetic category perception at birth. However, as we mentioned above, the results showing that newborns are able to discriminate between contrastive speech sounds are not very constraining for models. Results about the categorical perception of certain consonant contrasts provide stronger constraints on models and it would definitely be interesting to perform an evaluation based on these results. However, we wanted to start with a more comprehensive, albeit less direct, evaluation. Indeed, the categorical perception results only concern a limited set of phonetic contrasts and are typically obtained with synthesized stimuli, so that they would only allow evaluating models with respect to how they represent a very limited portion of the typical range of speech signal to which an infant is exposed. In contrast, the evaluation method we propose provides global measures of how well the different phonetic categories of a language can

be discriminated from each other based on the representation of speech signal by the model to be evaluated. We use specifically ABX-discriminability measures, which provide graded assessments of the discriminability of phonetic categories based on annotated speech recordings in the target language (cf. Section 2.1). There are many benefits in using ABX-discriminability measures, that we discussed at length in Chapter 1 and Chapter 2, but in particular, we can interpret them as providing predictions from the evaluated models regarding the expected performance of newborns in discriminating between phonetic categories (cf. Section 2.3.2). Unlike actual empirical measurements in newborns, these predictions are easily obtained for a large number of phonetic contrasts, with natural stimuli containing many sources of variability and take the form of graded measures of discriminability instead of binary assessments (*discriminable* vs *not discriminable*).

An obvious question is, what is the point of looking at such predictions in the absence of matching empirical evidence? The main idea is that it provides a convenient way to assess similarity and differences in how phonetic categories are represented by the different models at a functional level, beyond apparent differences in implementation. This can be used, in particular, as a basis for making modeling decisions in the absence of decisive empirical evidence or for designing new behavioral or cerebral imaging experiments. For example, if our goal is to use models of phonetic category perception at birth as a starting point for modeling phonetic category acquisition, there are many potentially interesting combinations of initial models and learning algorithms, too many for us to be able to test all of them. To restrict the number of initial models to be considered, we can apply simple *efficiency* and/or *simplicity* heuristics to our evaluation of these initial models. The idea of an *efficiency* heuristic would be to reject models that do not separate phonetic categories well enough. The idea of a *simplicity* heuristic would be to prefer simpler or more elegant models among models which separate phonetic categories equally well. Regarding the design of new behavioral or cerebral imaging experiments, our evaluation method might be used to devise experiments that can effectively decide between several alternative models. For example, one could look for the specific phonetic contrasts for which the predictions of the candidate models differ the most and focus the experimentation on these particular contrasts.

3.2 First step: motivating some candidate models from ASR

In the previous section, we introduced a methodology for designing and evaluating potential models of phonetic category perception at birth. The first step in this methodology consists in finding potential models by drawing inspiration from both models of human audition and models directly optimized for speech processing. In this section, we consider specifically a family of models suggested by ASR practice, which includes classical MFC and a version of PLP coefficients as special cases. Following our methodology, we investigate how they can be related to models of human audition and to optimized speech processing. Regarding the latter, the very fact that these models are used in ASR systems seems to implicitly suggest that they are optimized for speech processing. However, as we will see in Section 3.2.5, there is sometimes room for discussion, so we also try to make explicit in what sense the models can be considered useful for speech processing.

Let us now discuss how we chose the family of models considered. We could have used out of the box feature extraction packages, and directly compared them. A problem with this approach is that there are many possible variations in the detailed numerical implementation of speech features extraction methods and each feature extraction package incorporate design choice, specific parameters, and processing steps all integrated within a single black box. It is difficult to know which of these choices are critical and which are incidental for a given functional outcome, and in addition, it is very difficult to relate these black boxes with known properties human audition. Instead, our approach is to deconstruct these packages into a number of specific blocks with clearly identified functions, and then reassemble them in a large number of pipelines by selectively turning on and off each component and/or varying each parameter. In particular, to maximize the interpretability of our results, we computed all the representations under study using the same codebase, adapted from Dan Ellis' audio toolbox [155] (and from Sriram Ganapathy's code for FDLP coefficients [156], available at `http://www.clsp.jhu.edu/~sriram/research/fdlp/feat_extract.tar.gz`). We made sure that they only differed in the aspects explicitly mentioned in Section 3.2.1 below.

We end up with a whole family of possible processing pipelines (see Figure 3.1), which are introduced in Section 3.2.1. We group the different signal processing operations involved

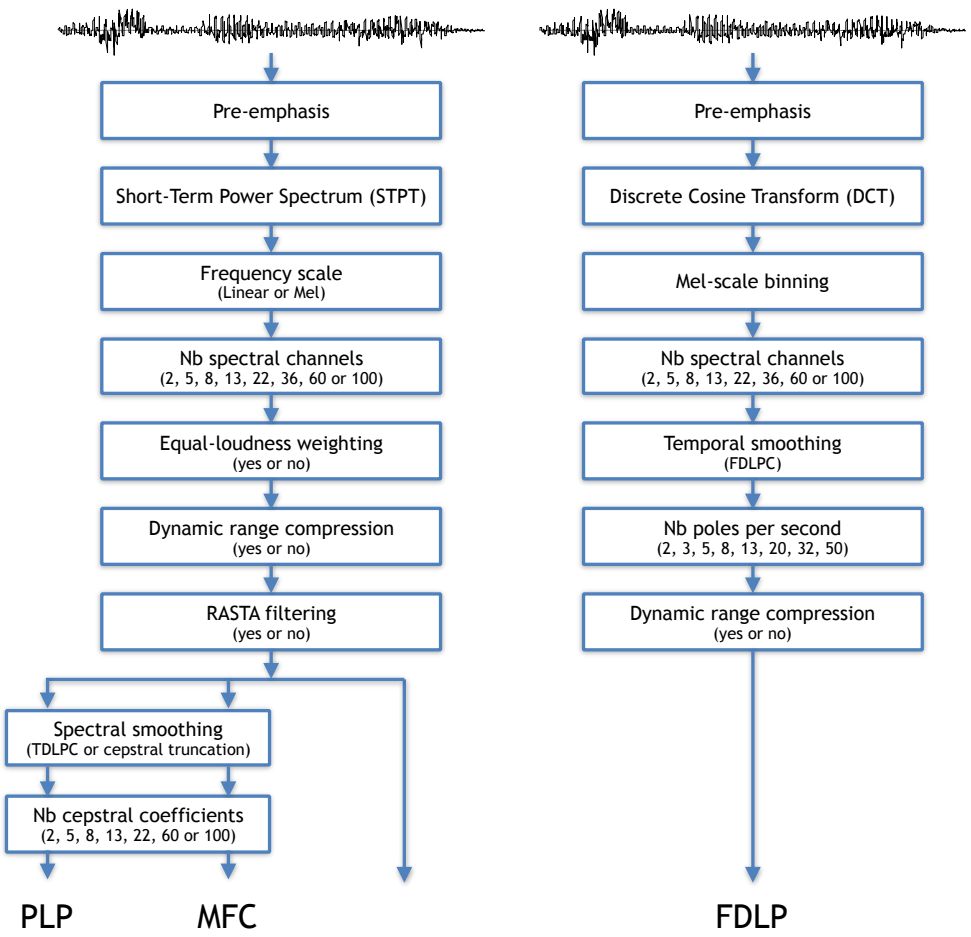


Figure 3.1: *Speech features extraction pipelines considered in this study. See Section 3.2.1 for a description of the different operations. The main pipeline is the one on the left of the figure. Classical MFC coefficients [18] are obtained with the following decisions: mel frequency scale (the number of spectral channels is not really standardized but is often taken between* 20 *and* 50*), no equal-loudness filtering, no dynamic range compression, no RASTA filtering and cepstral truncation (usually keeping* 13 *cepstral coefficients). A version of PLP coefficients, closer to [126] than to the original [112] (see Section 3.2.1 for details) is obtained with the following decisions: mel frequency scale (here too the number of spectral channels is not really standardized but is often taken between* 20 *and* 50*), no equal-loudness filtering, dynamic range compression, no RASTA filtering, Time-Domain Linear Predictive Coding (TDLPC) (usually with a* 5*-th order model, yielding* 6 *cepstral coefficients). The pipeline on the right of the figure is used exclusively to extract FDLP coefficients [156, 157], which we did not test as systematically as other features. This pipeline uses a different kind of mel-scale spectrogram than the others, which is obtained by applying Frequency-Domain Linear Predictive Coding (FDLPC) to sub-band spectrums obtained by binning a Discrete Cosine Transform (DCT) of the whole speech waveform according to a mel scale of frequencies. The most important difference between the two pipelines is that for MFC, respectively PLP, coefficients, cepstral truncation, respectively TDLPC, models spectro-temporal modulations along the frequency axis for each time slice, while for FDLP coefficients, FDLPC models spectro-temporal modulations along the time-axis for each frequency channel. The other differences between the two pipelines appear less fundamental, as we explain in Section 3.2.5.*

in these pipelines thematically to discuss how they relate to models of human audition and efficient speech processing. In Section 3.2.2, we discuss the *pre-emphasis* and *equal-loudness weighting* operations, which can be related to auditory processing at the level of the outer and middle ear. In Section 3.2.3, we discuss the *STPS*, *frequency rescaling* and *dynamic-range compression* operations, which can be related to auditory processing at the level of the inner ear. In Section 3.2.4, we discuss the *RASTA filtering* operation, which can be seen as a form of short-term adaptation process. Finally, in Section 3.2.5, we discuss the *cepstral truncation* and *time-domain LPC* operations, as well as *FDLP* coefficients, which can all be interpreted as different ways of statistically modeling spectro-temporal modulations in the speech signal. A rough summary of our discussion can be found in Table 3.1.

3.2.1 A classical family of speech features extraction methods

In this section, we introduce briefly the family of models considered. We study representations obtained at various stages in a speech processing pipeline (represented in Figure 3.1) leading among others to standard MFC [18] and a version of PLP [112, 126] coefficients. The first step in this pipeline, called *pre-emphasis* consists in applying a high-pass filter to the input signal. This was applied to all our pipelines. A Short-Term Power Spectrum (STPS) representation of the speech waveform is then obtained by applying a Fast Fourier Transform (FFT) to frames of 25ms duration taken every 10ms and taking the squared magnitude of the resulting complex Fourier coefficients. This was also applied to all of our pipelines. Next, one of 16 possible representations is derived by making a succession of 4 binary choices: use a linear or a Mel frequency scale; weight frequency channels according to human's equal-loudness contour or not; cubic root compress the dynamic range of frequency channels or not; apply log-domain RASTA filtering [158] to each frequency channel or not. For each representation, we test different resolutions for the frequency scale, using 2, 5, 8, 13, 22, 36, 60 or 100 frequency channels. To complete our study we apply Time-Domain Linear Predictive Coding (TDLP) to some of the 16 representations and re-estimate a cepstrum from the filter coefficients. In particular, we obtain a version of PLP coefficients, similar to the one recommended by [126], through the following path in the pipeline: *Mel scale/no equal-loudness/compression/no RASTA/TDLPC/cepstrum estimation*. The main differences with classical PLP coefficients [112] are that pre-emphasis is used instead of equal-

ASR Operation	**Human Audition**		**Function in Speech Processing**
	Physiology	*Psychology*	
Pre-emphasis/ Equal-loudness	Outer and middle ear	Loudness perception across frequency channels	Compensates for the spectral tilt in the glottal source
STPS filterbank	Inner ear	Auditory filters (frequency selectivity)	Reveals harmonic structure in the signal (formants)
Frequency rescaling	Inner ear	Perceptual distance for pitch (Mel) and Auditory filters bandwidth (Bark)	Stabilizes representation against small elastic deformations
Cubic-root compression of the dynamic range	Inner ear	Loudness perception within frequency channels	Enhances salient spectro-temporal features
STPS envelope	*Neural MTFs?*	*Special status of envelope in chimaeric sounds? Intelligibility of speech-shaped AM noise?*	*?*
RASTA filtering	*Neural short-term adaptation?*	*Context effects in auditory perception?*	Improves robustness to convolutional noise
Cepstral truncation (MFC)	*?*	*?*	Models spectrotemporal modulations along the frequency axis
Time-domain LPC (PLP)	*?*	*?*	Models spectrotemporal modulations along the frequency axis
Frequency domain LPC (FDLP)	*?*	*?*	Models spectrotemporal modulations along the time axis

Table 3.1: A rough summary of the associations we found between the signal processing operations in our ASR pipeline, aspects of human auditory physiology and psychology and the possible functions of these operations in speech signal processing. Entries with only an interrogation point indicate that we did not find any specific association. Other entries with an interrogation point indicate potential associations whose details remain very fuzzy. For the rest of the entries the association is clear, although this does not mean that it is perfect (for example, although certain operations can be clearly interpreted as modeling processing in the inner ear, this does not mean that they represent the best way to model processing in the inner ear). Note that we distinguished two parts in the STPS extraction step (see section 3.2.3 for more details): application of a filterbank (STPS filterbank) and envelope extraction (STPS envelope).

loudness filtering and that the frequency rescaling is performed using Mel-scaled triangular filters instead of Bark-scaled filters with a shape more closely inspired from physiological measurements. We also apply a cepstral truncation to some of the 16 representations, i.e. we apply a cepstral transform (log plus Discrete Cosine Transform (DCT)) and retain only the first n coefficients, where n is the desired number of cepstral coefficients). We obtain standard MFC coefficients

through the following path in the pipeline: *Mel scale/no equal-loudness/no compression/no RASTA/cepstral transform/cepstral truncation*. We study PLP and MFC coefficients based on 2, 5, 8, 13, 22, 60 and 100 cepstral coefficients (for PLP the number of cepstral coefficients can be shown to be the order of the autoregressive model plus one).

The results reported in this chapter were originally published in two separate studies [35, 96]. In the second study we performed some experiments with speech in noise and for these experiments all the cepstral representations are cast back in the spectral domain and an additional speech representation, FDLP coefficients [156, 157], is tested. Unlike the other representations, FDLP coefficients are not based on a time-frequency representation obtained through a Short-Term Fourier Transform but rather on a time-frequency representation obtained by taking the Inverse Discrete Cosine Transform (IDCT) of each bin in a Mel-binned Discrete Cosine Transform (DCT) of the whole sound. Frequency-Domain Linear Predictive Coding (FDLPC) is applied to this time-frequency representation to obtain the FDLP coefficients. It is not clear how fundamental the differences between the two types of time-frequency representations really are, but we defer further discussion of this issue to Section 3.2.5.

3.2.2 Motivating pre-emphasis and equal-loudness weighting

In this section, we consider how the *pre-emphasis* and the *equal-loudness weighting* steps relate to aspects of human audition and what sense they make as speech signal processing operations. We begin by considering the *pre-emphasis* step, which gives a higher weight to high frequency components by applying a simple high-pass filter to the signal ($y[n] = x[n] - 0.97x[n-1]$ for a signal x sampled at 16KHz).

The *pre-emphasis* step can be related to human audition through empirical observations that humans perceive sound energy in mid-range frequencies (typically between 500Hz and 5000Hz) better than in lower or higher frequencies. This is apparent, for example, in measures of auditory thresholds using pure tones at different frequencies: tones with mid-range frequency can be detected at a physical intensity at which tones with lower or higher frequency cannot be heard (see Figure 3.2, reproduced from [159]). Perhaps more directly relevant to speech signal, whose intensity is usually well above auditory thresholds, pure tones with mid-range frequency are perceived as comparable in perceived loudness to lower or higher frequency pure tones with

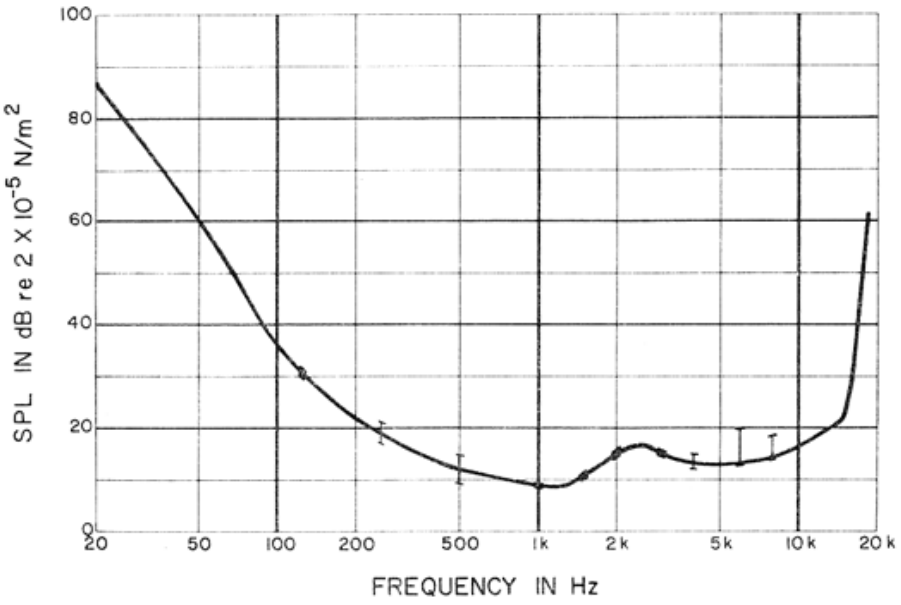


Figure 3.2: *Minimum audible pressure at the level of the eardrum for pure tones in humans as a function of sound frequency. Reproduced from [159].*

a higher physical intensity (See Figure 3.3, reproduced from [160]). From a physiological point of view, these effects are thought to result mostly from the action of the outer and middle ear, which act as a bandpass filter as they convey incoming sounds to the inner ear [163]. Note that all the empirical observations mentioned suggest the use of a bandpass filter whereas in our *pre-emphasis* step a high-pass filter is used, but the speech signal does not contain a lot of energy above 5000 Hz [164], so that these difference might have little effect in practice.

The *pre-emphasis* step also makes sense as a speech signal processing operation. Speech production can be described in terms of sound sources, glottal and/or fricative, being filtered by the vocal tract [165]. The energy in the spectrum of glottal pulses is known to decrease regularly with increasing frequencies, a property called the *spectral tilt* (See Figure 3.4, reproduced from [162]). This results in an imbalance in the long-term speech spectrum for which *pre-emphasis* is compensating.

In our ASR pipeline (Figure 3.1), after the next two steps -*STPS extraction* and *frequency rescaling*- we mention the possibility of applying an *equal-loudness weighting* to the frequency channels. This operation is conceptually very similar to *pre-emphasis*, the only difference being that the filter involved matches more closely the behavioral data from Figure 3.3. As such it would have been interesting to compare pipelines with *pre-emphasis* and without *equal-loudness*

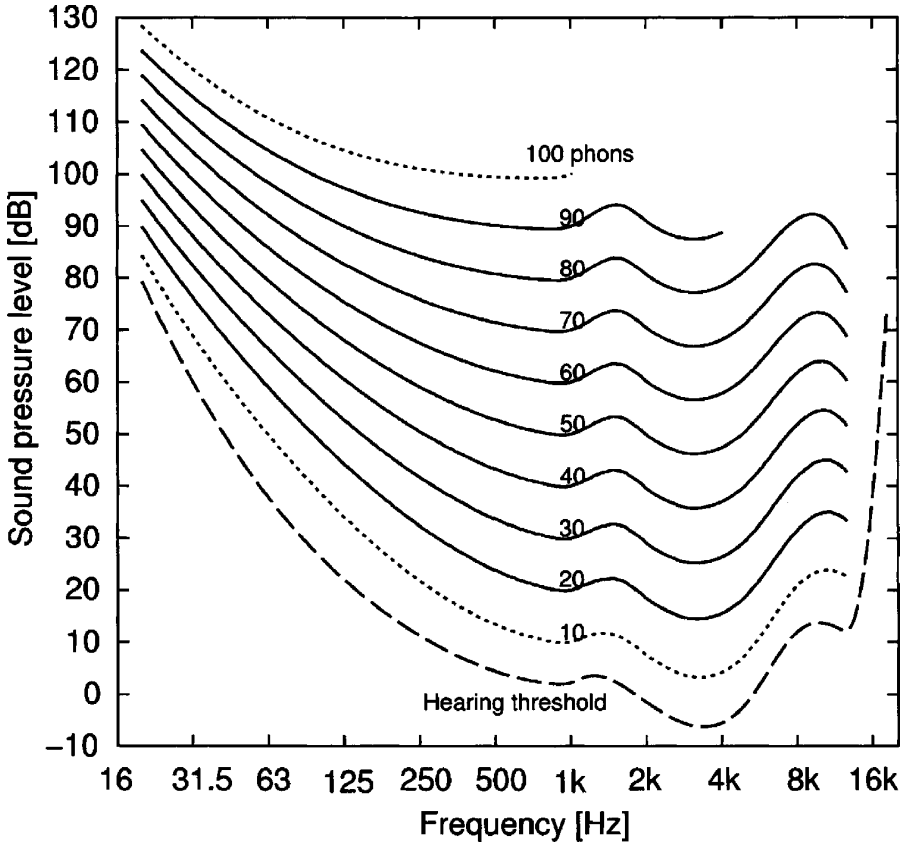


Figure 3.3: *Human equal-loudness-level contours for pure tones as a function of sound frequency for different loudness levels. The minimum audible pressure level is also indicated under the name* Hearing threshold. *Dotted lines indicate estimates based on limited empirical data. Reproduced from [160].*

weighting and vice-versa. Unfortunately, we realized this redundancy too late and we included *pre-emphasis* in all the pipelines that we tested., We therefore have to leave this comparison for future work.

3.2.3 Motivating STPS, frequency rescaling, and dynamic-range compression

The next steps in our pipeline consist in forming a time-frequency representation of the signal by estimating a Short-Term Power Spectrum (STPS), grouping the frequency bands according to either a Mel-scale or a linear scale, and applying a cubic root compression of the dynamic range or not. Let us first try and motivate their interest as speech signal processing operations and then discuss whether and how they can be related to models of human audition. STPS

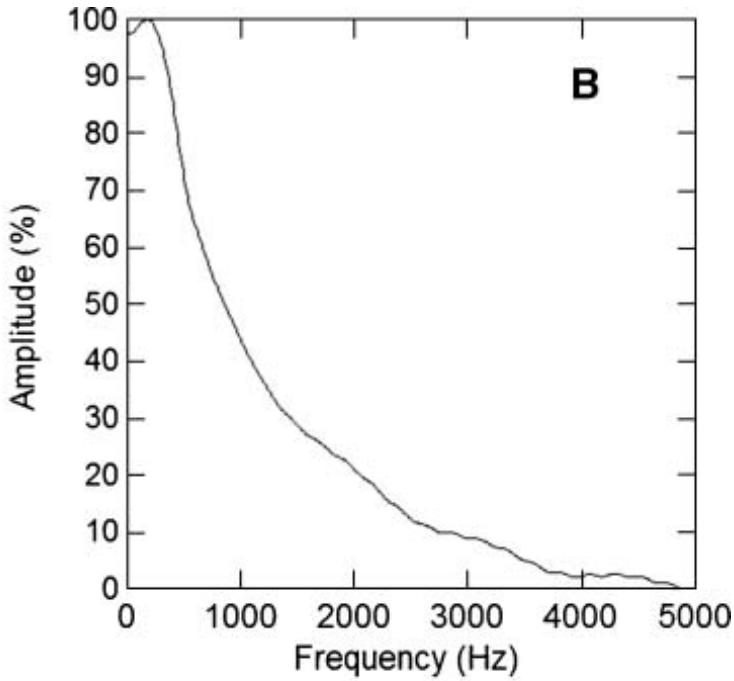


Figure 3.4: *Estimation of the spectrum of a single pulse from the glottal source based on the LF model of glottal source [161]. Reproduced from [162].*

extraction can be motivated as a speech signal processing operations by simple Source-Filter models of speech production [165, 166]. According to these models, the speech signal can be seen as resulting from the filtering of one or more source signals (glottal or fricative) by a resonant system (the vocal tract) whose properties are changing slowly relative to the frequency content of the source signals and are determined by the position of the principal articulators. This results in a signal containing a rich spectral content, including formants, fricative noise or pitch harmonics, which is revealed through STPS extraction. Regarding frequency rescaling on a Mel scale, its main effect is to produce frequency channels whose bandwidth increase proportionally to their center frequency. This operation has a clear interest in terms of signal processing: it can be shown to produce a representation whose high-frequency components do not suffer from the well-known instability of high-frequency components in the STPS (see for example [104], Section II.B). Finally, the cubic-root compression of the dynamic range of frequency channels can be interpreted as an operation enhancing salient features in the time-frequency representation, in the sense that it limits the impact on the distance between two sounds, of differences in intensity in time-frequency bins where both sounds have high intensity. This is useful, for example, to reduce the disruptive effects of background noise (see [167], Section III.F for an illustration of

this idea).

Now, to see what sense the STPS extraction, frequency rescaling and dynamic-range compression steps in our ASR pipeline can have within a model of human audition, we first need to interpret this sequence of three operations differently (the equal-loudness weighting operation, which is potentially applied between frequency rescaling and dynamic-range compression, is ignored in the following discussion, but it would be straightforward to include it). In Appendix B.1, we explain how the original sequence of three operations is approximately equivalent to a different sequence of three operations: convolution with a (real-valued) filterbank, compression of the dynamic range and envelope extraction. In the next paragraph, we show that the first two operations in this sequence approximately match simple phenomenological models of signal processing in the cochlea, which form an important building block of simple models of various auditory perceptual abilities. The last operation (envelope extraction) is less clearly related to known properties of human audition.

Incoming sounds are transmitted through the middle ear to the cochlea, an organ located within the inner ear that is responsible for the transduction of sound into neural activity. Consideration of the mechanical properties of the cochlea can lead to models of various complexity and degree of abstraction. The simplest models are purely linear but fail to capture important aspects of cochlear mechanics. One prominent aspect they fail to capture is the ability of the cochlea to encode the 120 dB dynamic range of sound intensity that human can perceive onto a limited 30-35 dB dynamic range of cochlear partition motion [168]. The simplest phenomenological models capturing this particular nonlinearity take the form of a Linear Translation Invariant (LTI) system (see for example [169]), followed by a static compressive non-linearity that adjusts the dynamic range (see Appendix B.2 for details). LTI systems are equivalent to filterbanks, so that the *convolution with a filterbank operation* in our (reinterpreted) pipeline can be seen as the LTI part of such a model, while the cubic-root compression of the dynamic range corresponds to the static non-linearity. When the reweighing of the frequency scale is done according to a Mel-scale, the filterbank involved in our ASR pipeline is similar in many respects to filterbanks directly fitted to empirical measurements in the cochlea, such as gammatone filterbanks [170]. In particular, it is constituted of a number of bandpass filters with different center frequencies regularly spread on a lin-log scale of frequencies and such that the bandwidth of the filters increases

proportionally to their center frequency. In contrast, when the reweighing of the frequency scale is done according to a linear scale, the center frequencies remain on a linear scale and all filters have the same bandwidth. Even when using a Mel-scale, some differences remain between the filterbank in our pipeline and more direct fits to cochlear measurements. In particular, in our ASR pipeline, the bandwidth of the filters depends on the number of filters, while this is not the case for cochlear models. This is because the frequency scale is reweighed using for each filter a triangular function which extends from the center frequency of the previous filter to the center frequency of the next filter. This could be an important difference, because we test models with different number of filters (2, 5, 8, 13, 22, 36, 60 or 100). Regarding the number of filters, motion in the cochlea is detected and transduced into neural activity by Inner Hair Cells (IHC) at approximately 3000 locations regularly spread along the length of the cochlear partition, suggesting a much higher number of filters for cochlear models than what is typically used in ASR. We did not test models with more than 100 channels however, because our results suggest that performance saturates for higher numbers of channels (see Figure 3.6). Regarding the static non-linearity, the cubic-root compression in our ASR pipeline is a good match to the observed compression of the dynamic range between an incoming sound power and the resulting motion of the cochlear partition, at least at typical intensities of speech signals (see [171] pp. 42-44).

We just saw that certain models in our ASR pipeline can be seen as simple phenomenological models of sound transduction in the auditory periphery. However, the auditory periphery is only the beginning of the auditory processing pathways in the brain, and it could be the case that further processing completely changes the nature of auditory representations. If it was the case, modeling the periphery would not be very relevant for our purpose, which is to model behavior, i.e. how the system function as a whole. Fortunately, the elements included in our models (convolution with a filterbank and dynamic-range compression) are known to have a significant impact on the operation of the whole auditory system. Indeed, bandpass auditory filters spread on a lin-log frequency scale and whose bandwidth is proportional to their center frequency are central to descriptions of auditory perception, as revealed by masking experiments (see [172], Chapter 3). The properties of the filters characterized behaviorally closely match those fitted to cochlear responses [173]. Frequency scales have been characterized behaviorally based on the

bandwidth of the filters estimated from masking experiments (Bark scale [174]) and based on the perceptual distance between sounds with different pitches (Mel scale [175]). In both cases, the scales are well described as lin-log scales, just like frequency scales estimated from cochlear measurements [176], although the different approaches lead to some differences in the frequency at which the transition between the linear part and the logarithmic part of the scale occur. In our ASR pipeline, we use the particular Mel-scale that was implemented in Dan Ellis's audio toolbox. The effect on the whole system of the cubic-root compression of the dynamic-range of frequency channels is revealed by perceptual judgments of the loudness of pure tones. Indeed, these judgments can be described by a power law with exponent .3 of the physical energy of the sounds (see [172], Chapter 4, Section 3),consistent with the idea that loudness is related to the total energy in a sound after dynamic-range compression. More generally, the convolution with a cochlear filterbank and dynamic-range compression operations are fundamental building blocks in many successful models of perceptual auditory abilities, such as, for example, models of loudness perception [177] or of the perception of auditory textures [178].

Thus far, we have found clear interpretations of the *convolution with a filterbank* and *dynamic-range compression* operations in our (reinterpreted) ASR pipeline in terms of modeling human audition. To finish this section, we turn to the *envelope extraction* step. This step has no clear analog in cochlear processing. Envelope extraction could occur at a later processing stage in the auditory pathways, but there does not appear to be definitive evidence for it, at least in the form it takes in our pipeline. Indeed while at the level of the IHC and Auditory Nerve (AN), there is a half-wave rectification and low-pass filtering of the signal, which can be interpreted as a form of envelope extraction, but the cutoff frequency is much higher than that used in STPS (on the order of 2-5 KHz versus 50Hz). Some form of demodulation operation is likely to occur at some point, as evidenced by the progressive decrease in the highest modulation frequencies to which neurons respond as one goes toward higher processing areas in the brain [179], as well as by experiments showing the special role of amplitude modulations in sound perception [180, 181], but there is no consensus on the precise form this demodulation can take. Many possible alternatives have been proposed [104, 182–185]. The particular form of demodulation used in our pipeline is a form of Hilbert envelope extraction which is known not to be appropriate to demodulate sounds with a harmonic carrier, like voiced speech [184]. Also, it is lossy as it only

exploits information from the retrieved modulator and alternatives have been proposed that can take advantage of information in the carrier as well [104]. In conclusion, it is not clear that the *envelope extraction* step in our ASR pipeline can be precisely related to human audition.

3.2.4 Motivating RASTA filtering

The next (optional) operation in our ASR pipeline is RASTA filtering in each frequency channel. As a speech processing operation, the main objective of RASTA filtering is to produce a speech representation more robust to some sources of external noise. Several versions of RASTA filtering have been proposed [158] and the one implemented in our pipeline correspond to band-pass filtering in the log-domain (the logarithm of an input signal is taken, a band-pass linear filter is then applied and the exponential of the result forms the output). The main idea of using the log domain is to remove from each channel slow multiplicative components due to convolutional noise from the source signal. This produces a representation more invariant to deformations introduced when the speech signal is transmitted through channels that do not have a uniform frequency response.

There is no direct evidence that such an operation is implemented in the brain, although there is plenty of evidence for complex short-term adaptation processes occurring at multiple processing stages. Already at the level of Inner Hair Cells and the Auditory Nerve, complex short-term adaptation processes are known to occur at several time-scales [186]. In terms of behavior, the auditory system can be shown to be especially sensitive to change in spectral patterns over time as demonstrated by a variety of context effects [187] and short-term adaptation processes are a likely candidate to explain at least some of these context effects (e.g. [188]).

3.2.5 Motivating cepstral truncation, TDLPC and FDLPC

The optional final step in our ASR pipeline consists in either applying a cepstral transform to the representation or first applying Time-Domain Linear Predictive Coding (TDLPC) to the representation and then applying a cepstral transform to the result. We also discuss in this section FDLP representations obtained by applying Frequency-Domain Linear Predictive Coding (FDLPC) to a time-frequency representation obtained in a different way than for the other pipelines. For all the operations considered in this section, we do not know of any clear

evidence that they are implemented in the brain, so we only discuss their interest in terms of speech signal processing. This does not mean that we have any particular reason to think that they could not be implemented in the brain.

The cepstral transform of a signal is obtained by taking the logarithm of this signal, applying a type-2 DCT transform on each spectral slice of the result and retaining only the N first coefficients. We tested values of N equal to 2, 5, 8, 13, 22, 60 and 100. Taking the logarithm of a spectral representation can be understood in terms of source/filter theories of speech recognition [165]: in the log-spectral domain, the convolution of a source signal with a filter is additive, making the two components more easily separable by linear models. The DCT transform is a kind of linear decomposition and based on the local correlation structure of the spectral representations considered (the correlation between two frequency channels is a decreasing function of the distance between their center frequency), it can be shown [189–191] to be asymptotically equivalent to applying a Principal Component Analysis (PCA), while being computationally much cheaper. This means in particular that the different coefficients in the resulting cepstral representation are approximately decorrelated, which is particularly appreciated in ASR because it allows to model them with diagonal-covariance Gaussian Mixture Models (GMM) which are computationally less expensive than full-covariance GMM.

The selection of a restricted number of coefficients in the cepstral domain results in the extraction of a smooth envelope along the frequency axis for each spectral slice of the representation (as can be seen by projecting the truncated representation back in the spectral domain). Applying TDLPC and recasting the resulting representation in the spectral domain can be seen as another way to extract an envelope for each spectral slice. Perhaps the main difference is that unlike envelope extraction based on truncating the DCT , TDLPC envelope extraction is motivated by an explicit model of speech production for voiced segments [166], according to which speech can be seen as a linear autoregressive process. In the case of cepstral truncation, the smoothness of the extracted envelope is controlled by the number of retained coefficients. In the case of TDLPC, it is controlled by the order of the autoregressive process. Finally, FDLPC can be interpreted as extracting an envelope from each frequency channels. So, while cepstral truncation and TDLPC can be seen as statistical models of speech modulations along the frequency axis for each spectral slice, FDLPC statistically models speech modulations along the

time axis for each frequency channel. The interest in modeling temporal and spectral modulations of speech can be understood intuitively in terms of speech production. The vocal tract acts as a filter whose resonances create a prominent spectral modulation structure in the signal, the formants, which are known to contain important linguistic information [165]. Speech is produced as a precisely timed sequence of articulatory gestures creating temporal modulations of speech which also contain important linguistic information [192].

3.3 Second step: testing how the models represent phonetic categories

In this section, we use ABX discriminability measures to compare the extent to which the different models introduced in the previous section are able to separate phonetic categories. The speech signal contains many sources of variability and the ABX discriminability framework we introduced in Chapters 1 and 2 allows us to study their impact separately by using different discrimination tasks. We look in particular at the effects of coarticulation, speaker identity and the presence of additive and convolutional noise. We also compute a complementary measure assessing the discriminability of talkers based on the representation of speech by the different models. We give details about the stimuli and tasks used and the measures performed in Section 3.3.1. The results are presented in Section 3.3.2.

3.3.1 Methods

3.3.1.1 Stimuli

We used stimuli from the AI-LSCP corpus [37] consisting in all possible Consonant-Vowel syllables of American English pronounced both in isolation and within a carrier sentence by 12 male and 8 female speakers with manual annotations of the beginning and end of the syllables. For the experiments without additive and convolutional noise, we used the isolated stimuli only (a total of 6839 stimuli). For the experiments with additive and convolutional noise, we used the sentence-embedded stimuli from 3 male and 3 female speakers for a total of 1709 stimuli. All recordings were sampled at 16KHz. We used three types of additive noise: white noise, babble-noise and car noise. For each sentence, white noise of the length of the sentence was

generated randomly and car and babble noise of the length of the sentence were sampled randomly from the Aurora-4 database. Before being added to the signal, the noises were scaled to obtain various Signal to Noise ratios (S/N), based only on the parts of both the signal and the noise corresponding to the CV. We tested seven different S/N: -20, -10, -5, 0, 5, 10 and 20 dB. We also performed a control with clean speech that we call the 60 dB S/N condition because clean speech typically has an S/N of approximately 60 dB. For convolutional noise, we used bidirectional filtering with two different order one Butterworth filters: a low-pass filter with a cutoff frequency of 100Hz and a high-pass filter with a cutoff frequency of 4000Hz.

3.3.1.2 Tasks

ABX tasks consist in presenting three stimuli A, B and X. We use Minimal-Pair ABX tasks (MP-ABX) where A and B differ only by some minimal phonemic contrast, and X is matched to either A or B. For the experiments without noise, we use three variants of the task, illustrated in Table 3.2. In the *Phoneme across Talker* task (PaT), A and B differ by one phoneme (either the vowel or the consonant) and are spoken by the same talker. X is spoken by a different talker but has the same phonemes as either A or B. It measures talker invariance in phoneme discrimination. In the *Phoneme across Context* task (PaC), A and B differ only by one phoneme and are spoken by the same talker. X, also spoken by the same talker, matches A or B in one phoneme and differs from both in the other phoneme, measuring context invariance in phoneme discrimination. In the *Talker across Phoneme* task (TaP), A and B are spoken by two different speakers and are phonemically identical. X is spoken by the same speaker as either A or B, but differs from them by one segment, enabling the measurement of talker discrimination.

Task	**A**	**B**	**X**	**Answer**
PaT	/ba/ sp1	/ga/ sp1	/ba/ sp2	A
PaC	/ba/ sp1	/ga/ sp1	/gu/ sp1	B
TaP	/ba/ sp1	/ba/ sp2	/ga/ sp1	A

Table 3.2: *Example of a possible choice of the A, B and X stimuli for each MP-ABX task not involving noise.* sp *stands for speaker,* PaT *for Phoneme across Talker,* PaC *for Phoneme across Context and* TaP *for Talker across Phoneme.*

For the experiments with noisy speech, we use two different tasks illustrated in Table 3.3. The first task is a *Phoneme across Talker within Noise* task (PaTwN): an across-talker MP-ABX

task, like the PaT task above, but within noise (i.e. the same kind of noise is applied to A, B, and X and, for additive noise, at the same SNR). The second task is used only for experiments with convolutional noise. It is a *Phoneme across Talker and Noise* task (PaTN): an across-talker MP-ABX task, like the PaT task, but also across noise in the sense that either A and B are noisy and X is noiseless or A and B are noiseless and X is noisy.

Task	**A**	**B**	**X**	**Answer**
PaTwN	/ba/ sp1 HF	/ga/ sp1 HF	/ba/ sp2 HF	A
PaTN	/ba/ sp1 HF	/ga/ sp1 HF	/ba/ sp2 LF	A

Table 3.3: *Example of a possible choice of the A, B and X stimuli for the two MP-ABX task used for speech in convolutional noise. For speech in additive noise only the first task is used and the noise condition is specified by both the type of noise and the SNR level.* sp *stands for speaker,* PaTwN *for Phoneme across Talker within Noise,* PaTN *for Phoneme across Context and Noise,* HF *for High-pass Filtered speech and* LF *for Low-passed Filtered speech.*

3.3.1.3 Model of the MP-ABX tasks

To perform the MP-ABX tasks on the basis of the speech representations a, b and x of the stimuli A, B and X, we follow the methodology described in Section 2.1, consisting in computing the DTW dissimilarities $d(a, x)$ and $d(b, x)$ between A, X and B, X on the basis of an underlying frame-based dissimilarity function. Then, the sign of $d(a, x) - d(b, x)$, is used to determine the response of the model (respectively B or A for a positive or negative sign) and an error rate is computed. The choice of the underlying frame-based metrics is important and may impact the results. Here, we follow the recommendation of [34] and use the cosine distance in all our tests.

3.3.1.4 Results analysis

The error rate score for a given MP-ABX task is defined as the average error rate over all the relevant triplets of stimuli A, B and X in the database. For the PaT and PaC tasks, we additionally compute average error rates over consonantal or vocalic contrasts. For the tasks without noise (PaT, PaC, TaP), we compute confidence intervals for the average error rates by bootstrap resampling across talkers. We also resample across talkers to perform significance tests when we test error rates differences.

3.3.2 Results

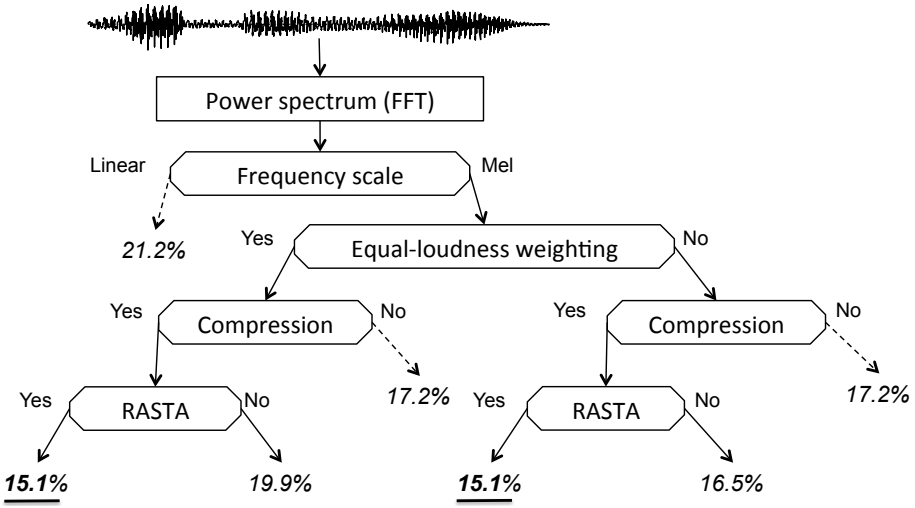


Figure 3.5: *First stages in processing pipelines for the computation of standard MFC and PLP coefficients. The MP-ABX error rate for the PaT minimal pair discrimination task is in italics. The best pipelines are shown with plain arrows, and the best scores are underlined. Parts of the pipeline not shown are indicated by dashed arrows and the best error rate achieved in each hidden part is indicated next to the arrow.*

We begin by looking at the effect of the number of spectral channels on the MP-ABX error rate (Figure 3.6). In Figure 3.6 (a), we look at a simple Mel-spectrum. In all tasks, we find an optimal number of channels, but it is not the same for the different tasks. The optimal number of channels is highest in the TaP task (36), intermediate in the PaC task (13) and lowest in the PaT task (8). This is a nice result since it has a simple interpretation: fine details of the spectrum contain a lot of speaker-specific information and little linguistic information, so that a coarser spectral resolution yields features more invariant to speaker change. However, the difference between the error rate for the optimal number of channels and the error rates for neighboring numbers of channels is small for all three tasks, so that it is worth asking how robust this result really is. We find that the difference in error rate between the optimal number of channels and neighboring numbers of channels is significant for the PaC and PaT tasks when resampling across talkers. This means that the precise optimal values for these tasks can be found robustly across talkers. We also find that it is not only for a Mel-spectrum representation that the optimal number of spectral channels is higher in the PaC task than in the PaT task. Indeed in Figure 3.6 (b) we see that it is also the case for the 8 representations from Figure 3.5

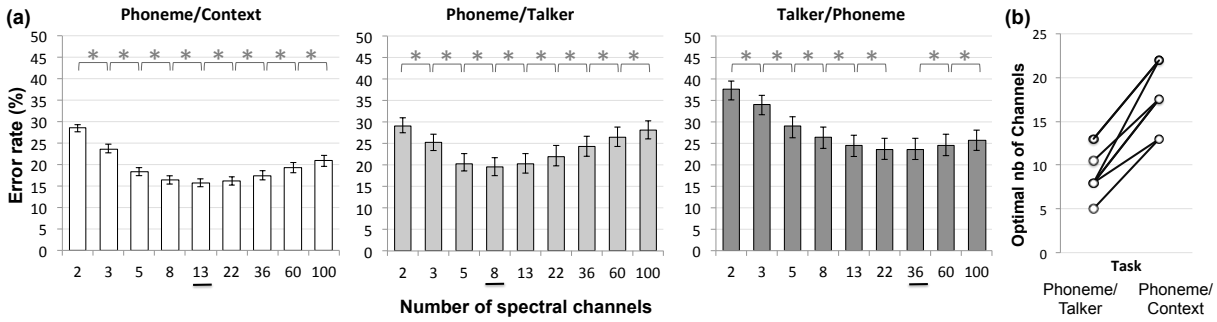


Figure 3.6: *(a) Average MP-ABX error rate in each task for a simple Mel spectrum with various number of spectral channels. Error bars represent 95% confidence intervals (sampled across talkers). The optimal number of channels is underlined and differences between error rates for adjacent number of channels that are significant at a level $\alpha = 1\%$ are indicated by a star. (b) Optimal numbers of channels in the PaC and PaT tasks for the 8 representations from figure 3.5 that are derived from a Mel-scale.*

that are derived from a Mel-spectrum. In the following, unless it is explicitly stated otherwise, we report error rates using the optimal number of spectral channels for each combination of task and speech representation.

The error rates for the different representations in the PaT task are reported in Figure 3.5. Using a Mel-scale appears clearly beneficial: the worst error rate for a representation using a Mel-scale (19.9%) is better than the best error rate for a representation using a linear scale (21.2%). The best representations are also consistently obtained when using cubic root compression and RASTA filtering, which yield improvements of 2.1% and 1.4% respectively of the error rate for the best representation. The other effect we observe is that equal-loudness filtering has a detrimental effect (3.4% increase in error rate) in the absence of RASTA filtering. This is likely to be due to the redundancy of equal-loudness filtering with the pre-emphasis operation that was applied to all the representations we tested.

The positive impact on phoneme discriminability of using a Mel-scale and applying cubic-root compression of the dynamic-range is confirmed by looking at results on the PaC task (Figure 3.7 (a) and (b)). Results on the TaP task (same figure), indicate that using a Mel-scale or a linear scale of frequency does not seem to affect the ability to discriminate speakers, but that cubic-root compression of the dynamic-range makes it easier. We also see (Figure 3.7 (c)) that equal-loudness weighting does not appear beneficial to phoneme discrimination, once again probably because of its redundancy with pre-emphasis.

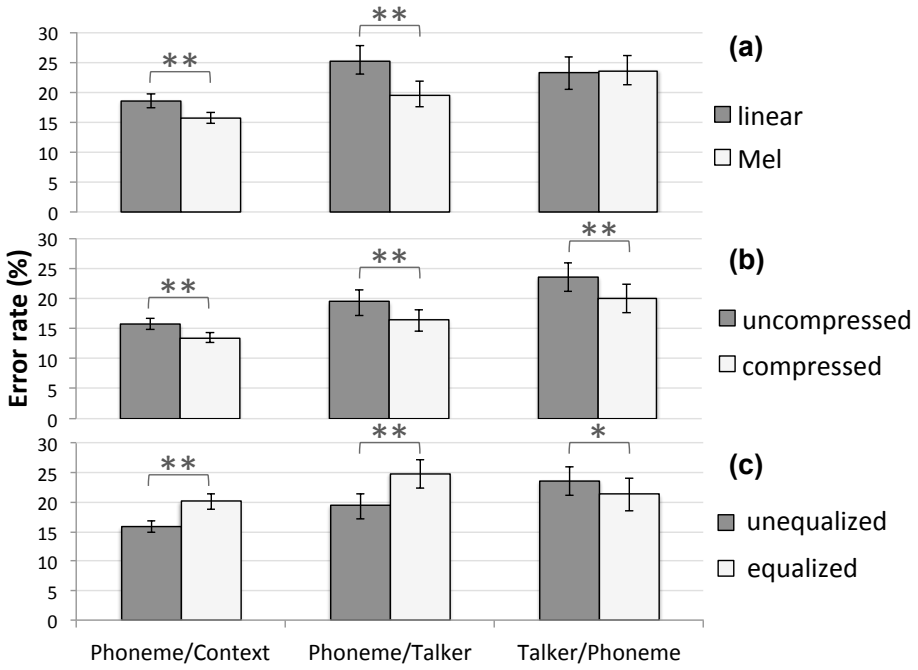


Figure 3.7: *MP-ABX error rate in the three tasks for (a) a simple Mel-scale or linear scale spectrum, (b) a Mel-scale spectrum with or without cubic root compression, (c) a Mel-scale spectrum with or without equal-loudness weighting. Error bars represent 95% confidence intervals (sampled across talkers). Differences significant at a level* $\alpha = 5\%$ *and alpha* $= 1\%$ *are indicated by one and two stars respectively.*

The effects of RASTA filtering applied to a cubic-root compressed Mel-spectrum in the three tasks are plotted in Figure 3.8(a). RASTA filtering appears to improve the discriminability of phonemes across talkers at the same time that it impairs discriminability of talkers across phonetic contexts. This can be interpreted as a form of speaker normalization. This interpretation is also supported by the absence of impact of RASTA filtering on the discriminability of phonemes across contexts. We uncover additional details on the coding properties of RASTA filtering by looking at error rates for consonants and vowels separately (Figure 3.8 (b)). RASTA filtering improves consonant coding and impairs vowels coding across both contexts and talkers. Moreover, while RASTA filtering improves consonant coding in both tasks by a comparable amount (3.7% and 3.4%) it impairs vowel coding by a lesser amount in the PaT task (1.8%) than in the PaC task (4.5%). All these results are coherent with the view of RASTA filtering as a form of short-term adaptation, enhancing transients in the signal that are useful for discriminating consonants and removing speaker-specific steady-state information, which is helpful in discriminating vowels within a given talker but less so across talkers.

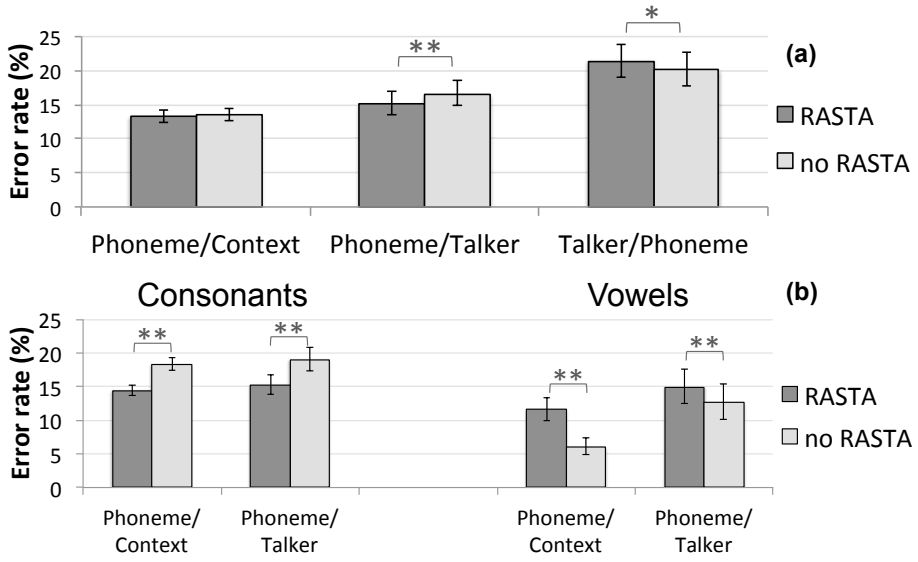


Figure 3.8: *(a) MP-ABX error rate in the three tasks for a cubic-root compressed Mel-scale spectrum with RASTA filtering or not. (b) Consonantal and vocalic MP-ABX error rates in the PaC and PaT tasks for the same representations. Error bars represent 95% confidence intervals (sampled across talkers). Differences significant at a level $\alpha = 5\%$ and alpha $= 1\%$ are indicated by one and two stars respectively.*

A potential caveat with our analysis of RASTA filtering, however, is that was done on the basis of isolated CV stimuli. This can be problematic because RASTA filtering, unlike the other signal processing operations we consider, averages activity at different instants in the time-frequency representation, which can be more problematic for stimuli that are not preceded and followed by silence. This is confirmed by a control performed with the sentence-embedded stimuli we use in the noisy phoneme discrimination tasks (in this control representations with a non-optimal number of 21 frequency channels were used): when comparing a Mel-spectrum representation with a RASTA-filtered Mel-spectrum representation in a PaT task using either isolated or sentence-embedded CV stimuli (Table 3.4), we find that the benefits of RASTA filtering are only present with isolated stimuli. RASTA filtering appears even a little detrimental with connected speech.

	Stimuli	
Pipeline	Isolated Speech	Connected Speech
Mel	25.2	25.9
RASTA Mel	21.9	27.0

Table 3.4: *Comparison of the MP-ABX error rates (in %) in the PaT task based on isolated or sentence-embedded CV stimuli for a Mel-spectrum or a RASTA-filtered Mel spectrum.*

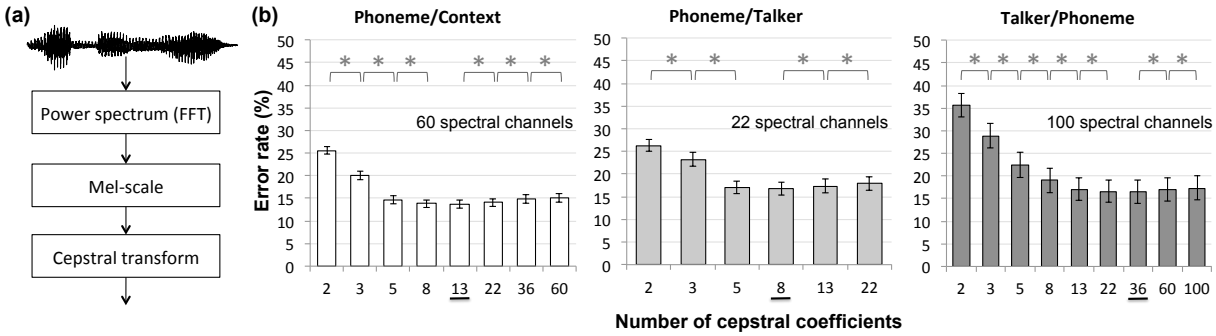


Figure 3.9: *(a) Processing steps for computing standard MFCC in our pipeline. (b) MP-ABX error rate in the three tasks for various numbers of cepstral coefficients, for classical MFCC. Error bars represent 95% confidence intervals (sampled across talkers). The optimal number of coefficients is underlined and differences between error rates for adjacent number of coefficients that are significant at a level $\alpha = 1\%$ are indicated by a star. The number of spectral channels was chosen for each task to optimize the minimal error rate in that task.*

There is one respect in which we still expect RASTA filtering to be beneficial even with connected speech: noise robustness, especially to convolutional noise. In Table 3.5, we compare a Mel spectrum with cubic-root compression of the dynamic-range with or without RASTA filtering (again using representations with a non-optimal number of 21 frequency channels), in tasks using connected speech under various noise conditions. We see, that although RASTA filtering is clearly detrimental in clean speech, it is beneficial in the presence of strong additive noise and convolutional noise. The positive impact of RASTA filtering is especially strong in the task *across* convolutional noise (15.1% improvement in error rate).

	Clean	**Add. Noise**		**Conv. Noise**	
Pipeline		-20 dB	Averaged	Within	Across
Compressed Mel	22.1	48.2	32.8	32.2	41.8
Compressed RASTA Mel	24.8	47.4	32.9	25.2	26.7

Table 3.5: *MP-ABX error rates (in %) in the PaTwN and PaTN tasks for a cubic-root compressed mel spectrum and a RASTA-filtered cubic-root compressed mel spectrum. All measures are computed based using connected speech. The* Clean *condition is a control based on clean speech (PaT task). The* Add. Noise *condition corresponds to results in the PaTwN task in the presence of additive noise. The results reported are averages of the results for the three types of noise (white-noise, car-noise and babble-noise). In the* -20 dB *sub-condition, results for the lowest S/N ratio tested are reported. In the* Averaged *sub-condition, results averaged over the seven S/N ratios tested are reported. The* Conv. Noise *condition corresponds to results in the PaTwN task (*Within *sub-condition) and PaTN task (*Across *sub-condition) in the presence of convolutional noise. The results reported as the averages of the results for high-frequency and low-frequency emphasis.*

We now investigate the final steps in our ASR pipeline, namely extraction of MFC, PLP or FDLP coefficients. First, we study the effect of the number of cepstral and spectral coefficients for standard MFCC. In contrast to previous representations, the number of spectral channels has a very small effect on the error rate for MFCC in the three tasks. Specifically, given a number of cepstral channels, changing the number of spectral channels did not change the error rate by more than 0.9% in the PaC task, 1.2% in the PaT task and 3% in the TaP task (note that the number of spectral channels need to be chosen larger than the number of cepstral channels for the representation to make sense). By contrast, the number of cepstral channels has a much bigger effect (Figure 3.9 (b)). Best results were obtained with 22 spectral channels in the Phoneme/Talker task, 60 in the Phoneme/Context task and 100 in the Talker/Phoneme task with respectively 13, 8 and 36 cepstral coefficients in striking accord with usual choices for these parameters in ASR. This is coherent with the idea already introduced that a coarser spectral resolution increases talker invariance, the difference being that for MFCC the spectral resolution is controlled by the number of cepstral coefficients rather than by the number of spectral channels. This is also true for PLP coefficients (where the number of cepstral coefficients corresponds to the order of the autoregressive model plus one).

Next, we compare MFC and PLP coefficients error rates on the three tasks (the optimal number of spectral and cepstral coefficients being chosen for each combination of task and representation). In Table 3.6, we compare results obtained using MFC and PLP coefficients cast in either the cepstral or spectral domain. We use a Mel spectrum with cubic-root compression of the dynamic range as a control. None of the representations obtained seems to improve over the compressed Mel spectrum baseline in either of the phoneme discrimination tasks. The results obtained are much worse for the MFC spectrum and similar or slightly worse for the PLP and MFC cepstrum and the PLP spectrum. The bad performance of the MFC spectrum is probably due to the absence of compression of the dynamic-range once the cepstral transform is reversed. Perhaps unexpectedly, the only area where the cepstral representation clearly improves over any spectral representations is talker discrimination. Cepstral-domain MFC coefficients, in particular, appear to support talker discrimination especially well.

Although MFC and PLP coefficients do not appear to improve much over a compressed mel-spectrum for phoneme discrimination in clean isolated speech, maybe things are different

Pipeline	**PaC**	**PaT**	**TaP**
Compressed mel spectrum	13.5	16.5	20.1
PLP cepstrum	13.6	16.6	17.4
MFC cepstrum	13.7	16.7	16.4
PLP spectrum	13.5	16.8	19.8
MFC spectrum	16.1	18.7	23.1

Table 3.6: *MP-ABX error rates (%) in the PaC, PaT and TaP tasks for PLP and MFC features expressed either in the spectral or the cepstral domain. Results for a compressed mel spectrum are also provided as a baseline.*

in connected speech in particular in the presence of noise. In Table 3.7 we compare the results obtained with MFC and PLP coefficients to results obtained with a compressed mel-spectrum and a RASTA-filtered compressed mel-spectrum in the additive noise discrimination task and the two convolutional discrimination task. We also include results obtained with FDLP coefficients. All the representations are cast back in the spectral domain and a cubic-root compression is applied to representations that do not include one by default (i.e. MFC and FDLP coefficients). All the representations are based on 21 spectral channels. The number of cepstral coefficients for MFCC and the order of the autoregressive models for PLP and FDLP coefficients are optimized.

	Noise				**Clean**
	Additive		Convolutional		
Pipeline			Within	Across	
Mel	36.2	41.7	43.0	25.9	
Compressed Mel	32.8		32.2	41.8	22.1
Compressed RASTA Mel	32.9		25.2	26.7	24.8
PLP	31.5		31.3	42.3	21.3
MFC	**31.0***		28.3	42.7	**20.6***
FDLP	**31.0***		**23.5***	**24.2***	23.4

Table 3.7: *MP-ABX error rates (in %): in the PaTwN task for additive noise (grand average across the three noise types and seven S/N ratios); in the PaTwN task (*Within *subcondition) and the PaTN task (*Across *subcondition) for convolutional noise (averaged over high frequency and low frequency emphasis); in the PaT task for clean speech. All results are obtained with connected speech stimuli. The best results for each condition are in bold font with an asterisk. All the tested representations are expressed in the spectral domain and cubic-root compressed.*

We consider first the results in clean speech. The PLP and MFC smoothed spectrum appear to improve over the compressed mel-spectrum baseline and the FDLP smoothed spectrum appears worse. The improvement of PLP and (compressed) MFC spectrum over compressed

mel-spectrum, which contrasts with what was observed with isolated speech (Table 3.6) could be explained by a specific advantage of PLP and MFC modeling when coding connected speech. However, it could also be due to the absence of optimization of the number of spectral channels, which, as we saw, does not impact MFC and PLP coefficients as much as the other representations. Further testing is needed to decide this issue.

In additive noise also, MFC and PLP coefficients appear to improve on compressed mel-spectrum with a comparable margin, but the interpretation of the result is again limited by the absence of optimization of the number of spectral channels. Interestingly, the (cubic-root compressed) FDLP-smoothed spectrum appears to perform as well as MFC coefficients and better than PLP coefficients on average in additive noise. This is remarkable because the FDLP-smoothed spectrum is expected to suffer from an unoptimized number of spectral channels as much as a compressed mel-spectrum (because the smoothing occurs along the time-axis in FDLP).

In convolutional noise, MFC and PLP coefficients do not appear to improve markedly over the compressed mel-spectrum and are much worse than a RASTA-filtered compressed mel-spectrum. Surprisingly, FDLP-smoothing appears to perform even better than RASTA-filtering for convolutional noise compensation, getting very close to clean speech performance in both the PaTwN and PaTN tasks.

3.4 Discussion

Let us begin with a summary of what was done in this study. We introduced a two-step approach to the design and evaluation of computational models of phonetic category perception at birth. In the first step, potential models are formulated based on two modeling principles: first, that generic auditory processes underlie speech perception at birth and second that speech perception at birth is efficient in extracting linguistic content from the speech signal (in a non-language-specific way). In the second step, the ability of the different models identified in the first step to separate phonetic categories is evaluated using ABX-discriminability measures. To illustrate our two-step approach, we focused on a family of speech feature extraction methods commonly used in ASR. We showed (first step) that most of the signal processing steps involved could be motivated in terms of efficient extraction of linguistic content from the speech signal and that

many could also be motivated in terms of modeling human audition. We then measured (second step) the impact of different processing steps on the discriminability of phonetic categories. We observed, in particular, that both frequency channels distributed on a lin-log scale, instead of a purely linear scale, and cubic-root compression of the dynamic-range of the frequency channels improve the discriminability of phonetic categories. We also observed that phonetic categories were better separated by representations with a rather coarse spectral resolution. The status of short-term adaptation in the different frequency channels using RASTA filtering is more ambiguous, because although we did find that it benefits the discriminability of phonetic categories in the presence of convolutional noise, it appears to have a detrimental effect in the case of clean connected speech. Regarding two classical signal modeling methods, cepstral truncation, as used in MFC features, and linear predictive coding in the time-domain, as used in PLP features, we found some evidence that they may benefit phonetic categories discriminability in connected speech, but additional controls with different numbers of spectral channels are needed before this can be confirmed. Linear predictive coding along the frequency axis, as used in FDLP features, appears clearly beneficial in the presence of additive or convolutional noise. It also appears somewhat detrimental in the case of clean speech, although additional controls with different numbers of spectral channels are also needed to confirm this result.

Overall, our results lead us to recommend as a model of phonetic category perception in infants, at this point, among the representation we tested, a simple pre-emphasized mel-scale power spectrum with cubic-root compression of the dynamic range. As we showed in Section 3.2.3, it can be interpreted as a simple phenomenological model of sound processing at the level of the ear. We used such a model for example in a study of the separation of phonetic categories in adult-directed speech versus infant-directed speech [115]. To go further, a number of control experiments mentioned above should allow clarifying the impact, or absence thereof, of cepstral truncation and time-domain and frequency-domain linear predictive coding on phonetic category separation. The impact of pre-emphasis and equal-loudness filtering should also be investigated. Beyond this, our discussion in Section 3.2 of the possible interpretations of the signal processing operations in our ASR pipeline in terms of human audition models and efficient speech processing, suggested many other interesting experimental questions. For example, we could test whether there is any functional benefit in using models for equal-loudness filtering that

match behavioral measurements more closely. We could also test filterbanks which match more closely behavioral and physiological measurements in humans, including with filter bandwidth independent of the number of filters. More generally, the potential functional benefits of more detailed phenomenological or physiological models of cochlear processing could be investigated. The question of the sub-band envelopes extraction step is also interesting. We could not find any definitive reason motivating envelope extraction as it is implemented in STPS extraction, be it in terms of modeling human audition or in terms of efficient speech processing. FDLPC, as used in FDLP coefficients, at least in a slightly different version than the one we tested [193], can be seen as an alternative way of extracting sub-bands envelope, but many other alternatives have also been proposed [104, 182–185]. Finally, there are plenty of other features extraction techniques in ASR (e.g. [194, 195]) and models of aspects of human audition (e.g. [196, 197]) from which inspiration might be drawn.

Additional evaluation tasks could also be used to bring more insight or reinforce existing results. For example, all the tasks were performed on the basis of read English stimuli for purely practical reasons, but there is no reason to think that infant perception at birth is particularly tuned to the phonetic categories of English or to read speech. Thus it would be interesting to try and reproduce the results based on stimuli in other languages and other registers, such as spontaneous or infant-directed speech. Also, because our database of stimuli included only one repetition of each CV stimuli by each speaker, we could only perform evaluation tasks with important sources of variability, such as a change in talker or phonetic context. It would be interesting to establish a baseline for phonetic categories discriminability in the absence of such sources of variability. Different measures than the average discriminability computed over all possible pairs of phonetic categories could also be used to gain more insight into the properties of different representations of speech. For example, discriminability scores for consonant and vowels can be computed separately, as we did when studying the effect of RASTA filtering, or, more generally, one could try to derive scores for broad phonetic categories or specific phonetic features.

Our approach has some limits. It allows to explore and organize the space of possible models and focus on those that are able to represent speech sounds adequately, but we are still likely to end up faced with several equally plausible and performant models. An extension of our work

(corresponding to a third step), could be to test these models more finely, for example on the basis of the results on categorical perception by newborns along certain speech continuums or by generating empirically testable predictions that would maximally discriminate between the identified alternatives.

In conclusion, in the absence of decisive empirical evidence, we were able to provide motivation and evaluation tools for models of phonetic category perception at birth, respectively by leveraging converging evidence from the speech science and the sciences of human audition, and by developing appropriate ABX discriminability measures. We hope, in particular, that this can prove to be a useful foundation for modeling the process of phonetic category acquisition during the first year of life.

Chapter 4

Modeling phonetic category perception in adults

Contents

4.1 Introduction

In this chapter, we ask to what extent Automatic Speech Recognition (ASR) systems, considered as computational models of speech processing in human adults, can account for the well-documented (see for instance [198] Chapter 1 or [199] Chapter 9) influence of the native language on how phonetic categories from foreign languages are perceived (for example native speakers of Japanese, a language where the /r/ and /l/ sounds from American English are not contrastive,

have a very hard time discriminating between these two sounds [200, 201]). We restrict ourselves to the question of the perception of foreign phonetic categories by strictly monolingual listeners, who have no or very little prior experience with speech in other languages. In particular, we do not discuss the question of native language (L1) influence on phonetic categories perception during second language (L2) learning.

We obtain a computational model of speech processing by speakers of a given language A, by training an ASR system with annotated speech in that language. This model is then presented with audio recordings of sounds from another language B, and the ABX-discriminability between phonetic categories in this other language is measured on the resulting output. This allows us to evaluate the model by comparing the patterns of discriminability it predicted with available empirical evidence regarding the confusions made by native speakers of language A between the phonetic categories of language B. We study the pattern of cross-linguistic phonetic category confusions in four languages : American English, Japanese, Mandarin and Vietnamese.

Let us motivate this study by reviewing how it relates to previous efforts in modeling cross-linguistic phonetic category perception by monolingual speakers. To discuss previous work, we need to introduce a distinction between models of general speech processing abilities and models of particular experimental tasks of interest. The effects of the native language on foreign phonetic categories perception are in most cases characterized empirically by having participants perform specific ABX or AX discrimination tasks. These tasks are used because they allow to nicely isolate and highlight certain specific properties of phonetic category perception. But the performance of participants in these tasks would be of little interest if it did not reflect more general properties of speech processing that are relevant for ecological speech processing tasks. It is thus important when designing computational models whose objective is to account for the behavior of participants in these tasks, to distinguish between parts of the model that correspond to general speech processing abilities that could be used in other contexts -which is what we are ultimately interested in- and parts of the model that correspond to using these abilities to perform the specific experimental task of interest.

The most prominent theoretical model of cross-linguistic phonetic category perception by monolingual speakers is the Perceptual Assimilation Model (PAM) [202–204]. PAM states that foreign speech is perceived in terms of the native language phonetic categories and that the

difficulty of perceiving a particular foreign phonetic contrast is determined by how the foreign phonetic categories involved map onto native categories. For example, if tokens from two different foreign categories are consistently mapped onto two different native categories, they are predicted to be easily discriminated. When tokens from different foreign categories are mapped onto the same native category, they are predicted to be harder to discriminate. Among foreign categories that map onto a same native category, PAM predicts that some are harder to discriminate than others based on *category goodness*: if tokens of one category are perceived as *good*, *typical* exemplars of a native category and tokens from another category are perceived as exemplars from the same native category, but *weird*, *unusual* ones, then the two categories are predicted to be easier to discriminate than if their tokens were perceived as equally good or equally bad exemplars of the native category. There are also some additional rules governing the case where one or both foreign categories involved are perceived as a non-linguistic signal (as in the case of clicks for native speakers of non-click languages), that we will not detail here. The main limitation of PAM is that it does not specify how to determine the mapping of foreign phonetic categories onto native phonetic categories. PAM postulates that speech is perceived in terms of the articulatory gestures needed to produce it, but it specifies neither the precise nature of this representation nor how it can be derived from the acoustic speech signal. PAM only provides a way to infer the relative discriminability of different phonetic contrasts based on a representation of speech sounds in terms of category labels and category goodness ratings. As such, PAM is best described as a *discrimination task model*, that can be applied to different *speech processing models*, where a *speech processing model* specifies a particular way of mapping foreign phonetic categories onto native phonetic categories.

Two other theoretical models of cross-linguistic phonetic category perception are often cited: the Speech Learning Model (SLM) [205–207] and the Native Language Magnet model (NLM) [208, 209]. Like PAM, these models do not specify how to determine the mapping of foreign phonetic categories onto native phonetic categories, but rely on certain aspects of this mapping to make predictions. SLM is concerned specifically with the plasticity of the mapping and as such is used mainly to make predictions regarding cross-linguistic phonetic category perception by long-time learners of a foreign language. Since we focus on cross-linguistic perception by naive listeners of a foreign language, we do not discuss SLM further here.

In NLM: *Experience is described as warping perception, producing a distortion that decreases perceptual sensitivity near category modes and increases perceptual sensitivity near the boundaries between categories* [209]. Phonetic segments at the *modes* of phonetic categories are supposed to correspond to phonetic segments that are judged good exemplars of their category by native listeners and phonetic segments near the *boundaries* are supposed to correspond to phonetic segments that are judged bad exemplars of their category by native listeners. NLM does not specify how to determine the category-goodness of a given sound in a given language, but given a set of sounds that are known to be good instances of a phonetic category in language A and bad instances of a phonetic category in language B, NLM predicts that they should be easier to discriminate by native speakers of language A than by native speakers of language B.

NLM is a simple phenomenological model, in that it offers no functional explanation for the phenomenon it describes. Feldman and colleagues [210] showed that this phenomenon emerges naturally in a system trying to infer the location of a stimulus in a perceptual acoustic space from a noisy observation, when this system operates in a statistically optimized way, taking into account the distribution of sounds in that acoustic space. This account holds under the assumption that the center regions and boundaries for the phonetic categories of the language considered align with modes and dips in the distribution of speech sounds observed by native speakers of that language in the acoustic space considered. A potential limit with that approach is that it assumes that speech processing in humans is optimized for estimating the location of a stimulus in a low-level acoustic space. In ecological conditions (by opposition to controlled laboratory experiments), however, differences between speech sounds that are not linguistically or para-linguistically meaningful can -and even should- be ignored. This suggests that it would be more natural to consider that speech processing in humans is optimized for classifying sounds into phonetic categories, i.e. for the *categorical perception* of phonetic categories rather than for estimating precisely their position in acoustic space. Of course, since the empirical observations from which NLM was formulated were obtained by having human participants perform experimental tasks involving within-phonetic category distinctions, it could be that the participants adapted specifically to these non-ecological tasks. Under this interpretation, the phenomenon described by NLM can be taken as evidence that native speakers of a language have internalized knowledge about the distribution of the speech sounds of this language in acoustic space.

However, there is no reason to think that this phenomenon occurs during speech processing in more ecological conditions.

Another functional interpretation of the phenomenon described by NLM is possible, based on the more natural idea that speech processing in humans is optimized for classifying sounds into phonetic categories. Bonnasse-Gahot and Nadal [211] nicely showed that a simple model of a population of neurons, whose objective is to classify a moderately noisy signal into discrete categories, is better off allocating its resources to encode fine distinctions near the category boundaries to the detriment of distinctions in more central regions of each category. Note that, contrary to the previous account, this one does not require the assumption that the center regions and boundaries for the phonetic categories align with modes and dips in the distribution of speech sounds. This means, that under this account the phenomenon described by NLM cannot be considered to be evidence that speakers internalize knowledge about the distribution of speech sounds in their native language. Note also, that the two functional accounts are compatible with each other, and both could be true at the same time and contribute independently to the effects observed empirically.

Although Bonnasse-Gahot and Nadal studied specifically a model of a population of neurons, the idea that resources should be allocated to encode fine distinctions near the category boundaries to the detriment of distinctions in the more central regions underlies virtually all existing approaches to classification in machine learning (e.g. Support Vector Machines [212]). This suggests that the behavior described by NLM will be naturally exhibited by any reasonable computational model of speech processing, and therefore that NLM provides no useful information for the design these models. Although it would be interesting to check this empirically, we did not attempt it in this study and we do not discuss NLM anymore in the following.

Let us now consider studies that examined the predictions of PAM using specific *speech processing models*. Different types of *speech processing models* have been investigated. Good results have been reported when the mapping between foreign and native categories was estimated empirically by having human subjects perform perceptual assimilation tasks [213]. This validates the use of PAM as a *discrimination task model*, but, since it uses humans as a proxy, it does not provide a computational model for the speech processing part. Predictions for the mapping between foreign and native categories based on an analysis of the phonology of the

two languages involved have also been tested, with disappointing results [214]. In particular, an abstract analysis at the level of phonemes appears doomed to fail, because phonetic and acoustic details in the stimuli, for example related to the phonetic and prosodic context, have been observed to influence the discriminability of segments. Predictions for the mapping between foreign and native categories based on the analysis of speech recordings, i.e. obtained from computational models of speech processing per se, have not received a lot of attention. We only found two studies looking at cross-linguistic phonetic category perception by monolingual speakers [214, 215]. Strange and colleagues [214] tried to predict cross-linguistic assimilation patterns of North German vowels into American English vowels by monolingual speakers of American English. Gong and colleagues [215] tried to predict cross-linguistic assimilation patterns of English consonants into Mandarin consonants by monolingual speakers of Mandarin. Strange et al. trained a Linear Discriminant Analysis (LDA) model to classify American English vowels based on manually-checked F1/F2/F3 formant frequencies and vocalic duration extracted from speech recordings. Gong et al. trained monophone HMM models used in ASR (they do not specify the nature of the emission probabilities of the models, presumably diagonal covariance Gaussians or Gaussian mixtures) to classify Mandarin consonants based on MFC coefficients extracted from speech recordings. The model in the study of Strange et al. was evaluated by looking at how North German vowels were classified and comparing with results from perceptual assimilation experiments with the same stimuli by native speakers of American English. Similarly, the model in the study of Gong et al. was evaluated by looking at how American English consonants were classified and comparing with results from perceptual assimilation experiments with the same stimuli by native speakers of Mandarin. The two studies report contrasting results. Gong et al. found a good match between their model's predictions and experimental results, while Strange et al. found many discrepancies.

There are three main differences between the research presented in this paper and previous approaches. First, we replace the PAM-based modeling of discrimination tasks by ABX discriminability measures. Second, we replace ad hoc speech processing models trained on very specific stimuli with general purpose ASR systems trained on natural continuous speech. Third, we expand the scope of the study to look at the full inventory of phonetic categories for four different languages. Let us motivate each of this points separately.

Using ABX discriminability measures has two main benefits. First, ABX discriminability measures can be directly related to the ABX discrimination tasks commonly used to study cross-linguistic phonetic category perception (cf. Section 2.3.1). Second, ABX discriminability measures can be seen as a generalization of predictions derived from PAM in the sense that, given a representation of speech segments consisting of a category label together with a category goodness rating, it is easy to provide a dissimilarity function such that PAM-based predictions and ABX discriminability measures computed with this dissimilarity function are compatible. The interest of the generalization is that unlike PAM-based predictions, that require categorical representations with category goodness ratings, ABX discriminability measures can be applied to any kind of representation for which a dissimilarity function can be provided.

Using general-purpose ASR systems has also several important benefits. First, it simply seems more natural to use as a model of speech processing in humans a general purpose system rather than a system only trained on isolated VCV stimuli with limited vocalic variability as in [215] or a system that can only recognize vowels and which uses speech features (F1/F2/F3 formants and duration) and an acoustic model (LDA) that are known not to be very performant for speech recognition purposes as in [214]. Second, the ability of the system to handle natural continuous speech, means that it can capitalize on the many existing corpora of annotated speech recordings. This allows training and testing systems in many languages and in a much more extensive manner. Third, using ASR systems means that the results are also of interest for the ASR community. Indeed, discrepancies found between ASR systems predictions and human behavior can provide insight into the shortcomings of ASR systems and inspiration for improving them.

By training and testing systems in four languages, we are able to evaluate the models in a much more comprehensive way. We are able to investigate both global effects in the perception of phonetic categories (for example phonetic contrasts of a language are globally harder to discriminate for non-native speakers than for native speakers [216]) and local effects (for example related to the perception of American English /r/-/l/ by Japanese listeners [200, 201], to the perception of stop categories based on VOT [217], or to the perception of tones by speakers of tonal and non-tonal languages [218]).

Let us mention a final motivation for this study: it paves the way for the study of models

of phonetic category *acquisition*. Because they are trained in a supervised way, using explicit annotations of speech recordings, ASR systems cannot be plausible models of how infants learn to perceive phonetic categories in a language-specific way. This does not mean that ASR systems cannot be good models of the language-specific perception of phonetic categories *at the end of the learning process*, i.e. in the adult, which is what we attempt to establish in this study. But, one might wonder why we did not test directly models of phonetic category acquisition instead. The reason is that, as we have seen in the introduction chapter of this thesis, the existing models of phonetic category acquisition suffer from strong methodological issues either in their design or in the way they were evaluated. In addition, the learning problem faced by these acquisition models is harder and has been much less studied both theoretically and practically than the supervised learning problem for ASR. As a consequence, the study of ASR systems appears as a safer first step to demonstrate the interest of the evaluation methods based on ABX discriminability that we developed. It also provides a useful baseline against which to compare future results obtained with models of phonetic category acquisition.

4.2 Methods

4.2.1 Corpora

To train and evaluate ASR models, 5 corpora of recorded speech in different languages were used:

1. A subset of the Wall Street Journal corpus (WSJ) [20], consisting of 143 hours of news article readings in American English by 338 speakers.

2. The Buckeye corpus (BUC) [99], consisting of 19 hours of spontaneous and rather casual conversational speech in American English by 40 speakers.

3. A subset of the Corpus of Spontaneous Japanese (CSJ) [124], consisting of 15 hours of spontaneous speech in Japanese by 75 speakers (in a more formal register than the Buckeye corpus, with one speaker relating an episode of his life to an audience).

4. The Global Phone Mandarin (GPM) corpus [219], consisting of 30 hours of news article readings in Mandarin by 132 speakers.

5. The Global Phone Vietnamese (GPV) corpus [220], consisting of 20 hours of news article readings in Vietnamese by 129 speakers.

Two corpora in American English were included to allow the separation of effects due to a change in language from effects due to other kinds of difference between corpora (e.g. difference in the properties of the recording microphones or in speech register).

Phonetic transcriptions were obtained from phonetic dictionaries and word-level transcriptions for the WSJ, BUC, GPM and GPV corpora. For the CSJ corpus, manual phonetic transcriptions were used. For all corpora, timestamps for the phonetic transcriptions were obtained by forced-alignment using a GMM-HMM ASR system similar to those described in the next section, but trained on the whole corpus.

4.2.2 ASR Models

We restrict our investigation to technologically-mature systems based on GMM-HMM architectures (by opposition to the more recent, but less mature, DNN-based systems). Each corpus was randomly split into a training and a test set, each containing an equal number of speakers. Then, a GMM-HMM ASR model with speaker adaptation and phonetic-context- and word-position-dependent phone models was trained with the Kaldi toolkit [221] on the training set of each corpora. The same Kaldi recipe was used to train all models (see `https://github.com/bootphon/abkhazia/blob/master/abkhazia/kaldi/kaldi_templates/train_and_decode.sh`). All models were also trained with the same parameters and input features. Phonetic-context-dependent phone models were obtained using triphone states with bottom-up clustering of the states to obtain a decision-tree with a total of 2500 leaves, involving a total of 15000 Gaussians. Input features consisted of 13 MFCC coefficients plus 3 pitch-related features [222] and their delta and delta-deltas coefficients. Pitch features were included because tone is contrastive in Mandarin and Vietnamese (i.e. there are words differing only by their tone in these languages).

The WER on the test set for each of the resulting models, using a word-level bigram language-model estimated from the training set, is reported in Table 4.1. The best performance is obtained on the WSJ corpus and the worse on the BUC corpus. The GPV, CSJ and GPM corpus are intermediate. The Kaldi recipe used was originally adapted from a recipe optimized for the WSJ corpus, which might explain the much better performance obtained on the WSJ. More

generally, many aspects of ASR technology have been developed for American English and might be somewhat overfitted to this language. The bad performance on the BUC corpus can be attributed to the speech register for this corpus, which is spontaneous and much more casual than for the other corpora.

Corpus	**WER**
WSJ	8.5%
BUC	48%
CSJ	30%
GPM	31%
GPV	23.5%

Table 4.1: Word-Error-Rates obtained for the ASR systems trained on each corpus

GMM-HMM models can provide an output in many different formats, such as word- or phone-level transcriptions, word- or phone-level n-best lists or lattices or frame-by-frame phone-level posteriorgrams of various kinds. We consider only Viterbi-smoothed phone-level posteriorgrams, which are more informative than phone-level transcriptions and are easily obtained from Kaldi models. A drawback is that these posteriorgrams, unlike raw acoustic likelihoods or Baum-Welch-smoothed phone-level posteriorgrams have no simple probabilistic interpretation. The posteriorgrams are obtained using a phone-level bigram language model estimated on the training set of each corpus. We do not claim that this choice of output representation is necessarily optimal and it would be interesting to test other possibilities.

4.2.3 ABX Evaluation

The test set of each corpus is decoded with each of the 5 ASR models to yield Viterbi posteriorgrams and the input features are added as a control to yield 6 different representations for each of the 5 corpora. For each corpus, a minimal-pair ABX task (see Section 2.1) ON phonetic category BY talker, previous phone and following phone (see Section 1.4) is compiled on the test set and used to measure the ABX-discriminability between phonetic categories for the 6 representations of that corpus. Since the number of word minimal-pairs is limited, we use single-phone minimal pairs in the ABX tasks (see Section 2.1), which means that the tasks are similar to the *within context* task of Section 2.4.2.1. An ABX triplet in such a task could be for example:

A	B	X
$/\mathrm{i}/^{T_1}_{b\_t}$	$/\mathrm{u}/^{T_1}_{b\_t}$	$/\mathrm{i}/^{T_1}_{b\_t}$

where ${}_{b\_t}$ indicates a segment preceded by a /b/ and followed by /t/ and ^{T_1} indicates a segment pronounced by speaker T_1. For each phonetic contrast, we compute a summary ABX score as follows. We start from ABX discriminability measures for each combination of talker, preceding context and following context for this contrast. First, we average out the talkers to obtain scores for each combination of preceding context and following context for this phonetic contrast, and then we average over the phonetic contexts, yielding a single score for each phonetic contrast. Depending on the experiments, we use directly the scores obtained for individual phonetic contrasts or we average these scores for all vocalic contrasts or for all consonantic contrasts to obtain summary measures of vowel, resp. consonant discriminability.

Dissimilarities between sequences of posteriorgrams are computed using Dynamic Time Warping (DTW) [19] based on a frame-to-frame symmetric Kullback-Leibler divergence [28]. Note that, contrary to what is done in PAM, dissimilarities are here based on more information than just category labels or category-goodness of the most likely categories. For the input representation control, we compute dissimilarities using DTW based on a frame-to-frame cosine distance. We do not claim that these choices of dissimilarity functions are necessarily optimal and it would be interesting to test other possibilities.

4.2.4 Analyses

We test whether ASR models can account for various effects in cross-linguistic perception of phonetic categories. We study both global effects, that involve all possible phonetic contrasts and allow for a very quantitative evaluation of the models, and more local effects, that involve specific phonetic contrasts and for which extensive empirical data is often available.

We compute two types of global measures based on the representations of a given corpus by the different models. We compute the average error rate of each model in discriminating the phonetic contrasts of the corpus. We also compute the similarity between the patterns of error of the different models. The two kinds of measures are meant to be completely independent: one measures to what extent a given model is able to separate the phonetic categories of a given

corpus, the other measures to what extent the phonetic categories that are hard and those that are easy are the same for two different models, independently of the absolute level of performance of these models. The independence is formally guaranteed by measuring the similarity between error patterns with the cosine similarity, which is invariant to rescaling of its arguments. In particular, if two error patterns are just globally rescaled version of each other they will be judged perfectly similar (cosine similarity equal to one).

Phonetic contrasts of a language are found to be globally harder to discriminate for non-native listeners than for native listeners (see e.g. [216]). Measures of average error rate are used to test whether models can account for this effect: we should find that phonetic contrasts of a language are globally harder to discriminate for models trained on a different language (mismatched-language condition) than for models trained on the same language (matched-language condition).

Although phonetic contrasts are globally harder to discriminate for non-native listeners than for native listeners, it is well-established that all contrasts are not affected to the same extent. Moreover, the way in which the contrasts are differentially affected is known to be largely determined by the native language of the non-native listeners [198]. Measures of similarity between error patterns are used to test whether models can account for this effect: we should find that the patterns of confusion for the representations obtained from the two American English ASR models are more similar to each other than to representations obtained from other ASR models, independently of the corpus on which the confusions are measured.

We also investigate five different local effects chosen. First, the perception of American English /r/-/l/ by Japanese listeners is known to be very poor [200, 201], so that we expect a model trained on CSJ to be much worse at discriminating this contrast than any model trained on American English. Second, Mandarin and Japanese stops involve different VOT categories but similar places of articulation [223]. It has also been reported in a conference presentation [224] that VOT contrasts of Japanese stops are hard to perceive for Mandarin native speakers, although no published data is available to the best of our knowledge. Thus, we expect Japanese stop contrasts to be comparatively easier to discriminate based on the GPM representation when they differ in place than when they differ in VOT. For the same reason, we expect Mandarin stop contrasts to be comparatively easier to discriminate based on the CSJ representation when they differ in place than when they differ in VOT. Third, Mandarin and

Vietnamese are tonal languages, while Japanese has only pitch accent and English only lexical stress. It is thus interesting to look at the discriminability of Mandarin and Vietnamese tones by the various models. Previous experience with a tonal language has been reported to be beneficial or detrimental for cross-language tone perception depending on the cases [225, 226]. Since, to the best of our knowledge no empirical data is available for our particular sample of languages, we are unfortunately not able to make any clear predictions. Fourth, vowel contrasts of Japanese differing only in duration have been reported to be hard to discriminate for native speakers of American English [227, 228]. Accordingly, we expect the WSJ and BUC models to be much worse than the CSJ model in discriminating these contrasts. Fifth and last, discriminability measures in ABX tasks have been obtained empirically for a few vocalic contrasts of American English in both Japanese and Mandarin native speakers [213, 229]. We expect the pattern of discriminability observed empirically in Japanese -respectively Mandarin- native speakers to be closer to that of the CSJ -respectively GPM- model than to any other model.

4.3 Results

4.3.1 Global Effects

The average ABX-discriminability for consonant contrasts (Figure 4.1a) and vowel contrasts (Figure 4.1b), show that in all cases a model trained on a corpus separates more easily phonetic categories in speech taken from that corpus (*matched corpus* condition) than in speech taken from a different corpus (*mismatched corpus* condition). We also see that the two models trained on a corpus of American English (WSJ and BUC) separate phonetic categories in that language better than phonetic categories from other languages, even when they are tested on a different corpus than the one on which they were trained. This means that the differences observed cannot be explained by low-level differences in the corpora, such as the type of microphone used to record the signal for example. This is all the more interesting since the WSJ corpus is actually more similar in many respects to the other corpora, in particular the GPM and GPV corpora, than to the BUC corpus. For example, the speech register and the topics are similar in the WSJ, GPM and GPV corpora (news articles readings) and different from the speech register and topics in the BUC corpus (spontaneous, and often quite casual, dialog with an interviewer

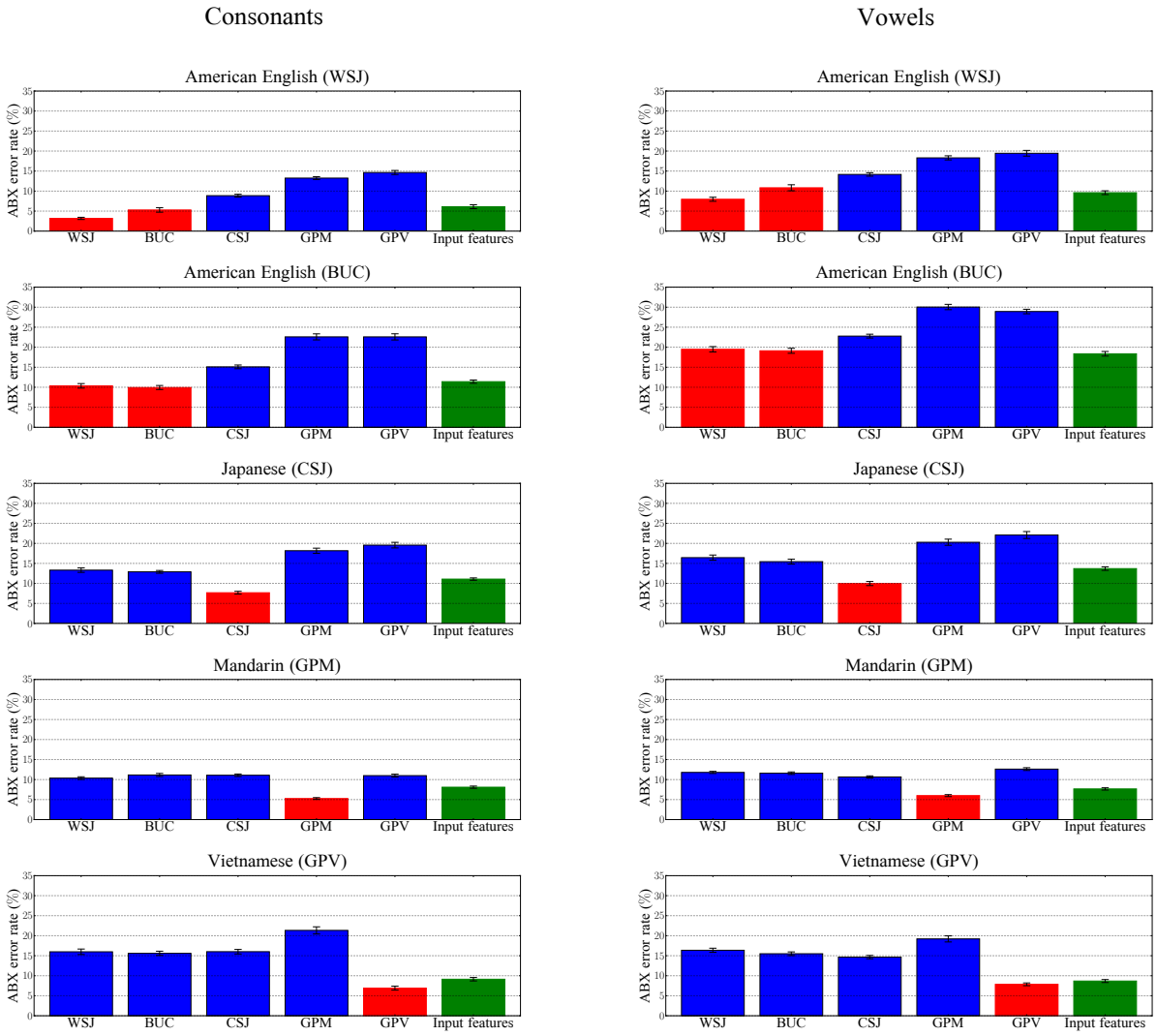
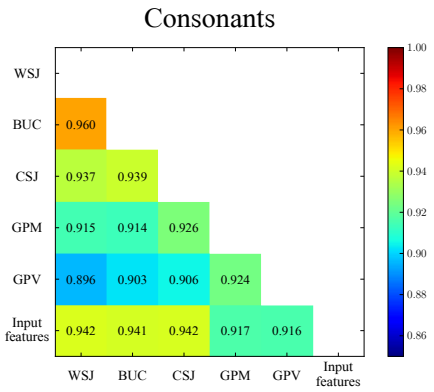


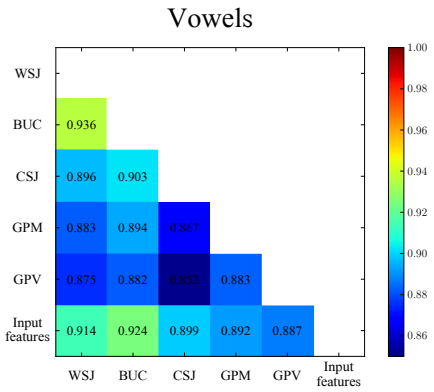
Figure 4.1: (a) Average ABX-discriminability of consonant contrasts. Columns correspond to the corpora on which the models were trained and lines correspond to the corpora on which they were tested. Red bars are used for the *matching language* conditions, i.e. where a model was tested on a corpus in the same language as the corpus on which it was trained. Blue bars are used for the *mismatched language* conditions, where a model was tested on a corpus in a different language than the corpus on which it was trained. Green bars are used for the baseline obtained by using as a representation the input features (MFCC plus pitch features) common to all models. Error bars indicates mean plus and minus one standard deviation and were obtained by bootstrap resampling of the scores at the level of the speakers. (b) As in (a), but for vowel contrasts.

on everyday topics) and in the CSJ corpus (relations, not read but somewhat prepared, of a memorable episode of their life by speakers in front of a small audience). These differences in register and topics only appear to have a global effect on the results: all models perform better

on the WSJ, GPM and GPV corpora than on the BUC corpus and the performance on the CSJ corpus lies somewhere between these two extremes. Another interesting observation is that training a model on a different language appears to render phonetic categories less separable than they were in the initial representation. Training on the same language yields results that improve over the initial representations in most cases, excepted for the vowels of the BUC corpus where there is even a slight worsening.



(a) For each pair of models: average over the different corpora of the correlation between the normalized ABX error rates for each pair of consonants.



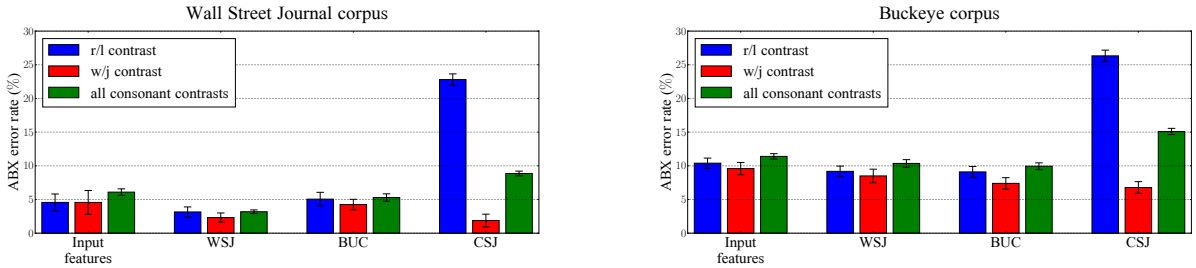
(b) For each pair of models: average over the different corpora of the correlation between the normalized ABX error rates for each pair of vowels.

We established that the different models yield language-specific representations, where phonetic categories of the language on which they were trained are easier to discriminate than phonetic categories of other languages. Next, we ask whether the difference in average discriminability between models in matched-language and mismatched-language conditions results from a global rescaling of the discriminability of each phonetic contrasts or whether, as expected from a model of phonetic category perception in human, certain contrasts are more impacted than others depending on the specificities of the training and test languages. For each pair of models, the cosine similarity between patterns of confusions is computed separately for confusions between consonants and between vowels for each corpus. The cosine similarities averaged over the five different corpora for each pair of models are plotted in Figure 4.2a and 4.2b.

The patterns of confusion appear quite similar for all pairs of models (minimum similarity: .852 for vowels and .896 for consonants). This can be interpreted as meaning that there are some intrinsically easier and harder contrasts irrespective of any model training. Furthermore, the similarity of any model with the input features is good, in many cases better than between that

model and models trained on other languages (minimum similarity with input features: .887 for vowels and .916 for consonants), suggesting that simple acoustic differences can account to some extent for these intrinsically easier and harder contrasts. The most interesting result is that for both consonants and vowels the error patterns of the BUC and WSJ models are more similar to each other than to any other model (and no other pair of models has more similar error patterns than this one). This shows that the patterns of errors obtained are language-specific. We also see that the GPM and GPV models have the error patterns that are the least similar to all the others and that the error patterns of the CSJ model are more similar to those of the WSJ and BUC models than to those of the GPM and GPV models.

4.3.2 Local Effects



(a) (b)

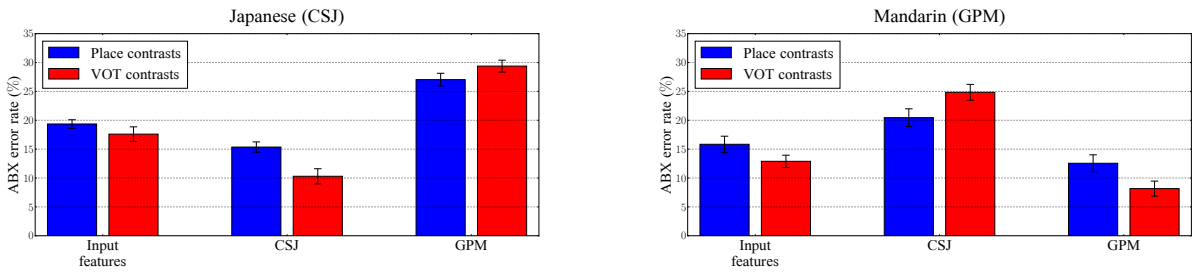
Figure 4.2: Average ABX error-rates obtained on the WSJ corpus and on the BUC corpus with four different representations (input features, WSJ, BUC, CSJ) for the /r/-/l/ contrast and two controls. Error bars indicates mean plus and minus one standard deviation and were obtained by bootstrap resampling of the scores at the level of the speakers.

Up to this point, we established that the overall performance of the different models, as well as their patterns of confusion between phonetic categories, were specific to the language used to train them. To further investigate the potential of these representations as models of phonetic category perception in humans, we now look at more specific effects that have been studied directly in humans. First, we consider the discriminability of the /r/-/l/ contrast of American English by models trained on American English and on Japanese. The prediction here is that this distinction should be much harder to make for the Japanese-trained model than for models trained on American English [200, 201]. In Figure 4.2, we plotted the ABX error-rates for the /r/-/l/ contrast obtained on the WSJ corpus and on the BUC corpus. We also plotted as controls the error-rate for /w/-/j/, another liquid contrast, and the average error-rate for all

consonant contrasts. We see that the discriminability of phonetic categories is globally more difficult in the BUC corpus than in the WSJ corpus, but that the same pattern is observed for both corpora, confirming that we are observing language-specific effects (and not corpus-specific effects). Looking at the input features baseline, we see that both liquid contrasts are slightly above the average discriminability of consonant contrasts with the /w/-/j/ contrast perhaps slightly easier than the /r/-/l/ contrast. When we look at the matching-language conditions, we see that in all cases the discriminability either improves or at least remains similar when compared to the input features discriminability. In the mismatched-language condition (i.e. for the CSJ model), we see that while the /w/-/j/ contrast becomes even more discriminable than for models trained on American English, the /r/-/l/ contrast becomes much much more difficult to discriminate. The extent of the degradation in the discriminability of the /r/-/l/ contrast is underlined by the comparing it with the degradation in discriminability averaged over all consonants which exists but is much smaller.

	Japanese			**Mandarin**		
Place	Labial	Alveolar	Velar	Labial	Alveolar	Velar
VOT						
Voiced	b	d	g			
Tenuis	p	t	k	p	t	k
Aspirated				p^h	t^h	k^h

Table 4.2: Stops in Japanese and Mandarin



(a) (b)

Figure 4.3: Average ABX error-rates obtained on the CSJ corpus and on the GPM corpus with three different representations (input features, CSJ, GPM) for place and VOT contrasts. Error bars indicates mean plus and minus one standard deviation and were obtained by bootstrap resampling of the scores at the level of the speakers.

Next, we look at another expected effect related to Voice-Onset-Time (VOT) distribution in

stop consonants. We contrast stop consonants in Mandarin and Japanese because while they have similar place of articulation they systematically differ in VOT (see Table 4.2) [223]. This allow to make predictions regarding the relative discriminability of VOT and place contrasts: using a model trained in a mismatched-language condition is expected to affect VOT contrasts (/p/-/b/, /t/-/d/, /k/-/g/ in Japanese and /p^h/-/p/, /t^h/-/t/, /k^h/-/k/ in Mandarin) more than place contrasts (/p/-/t/, /p/-/k/, /t/-/k/, /b/-/d/, /b/-/g/, /d/-/g/ in Japanese and /p^h/-/t^h/, /p^h/-/k^h/, /t^h/-/k^h/, /p/-/t/, /p/-/k/, /t/-/k/ in Mandarin). The results are plotted in Figure 4.3. For both languages, VOT contrasts appear easier to discriminate in the input features than place contrasts. In the matched language conditions there is a global improvement, both VOT and place contrasts become easier to discriminate than in the input features, but the pattern remains the same: VOT contrasts are easier than place contrasts. In the mismatched language conditions however, in line with the predictions, the pattern changes. There is a global decrease in the discriminability, both VOT and place contrasts become harder to discriminate than in the input features, but in addition, VOT contrasts become harder than place contrasts.

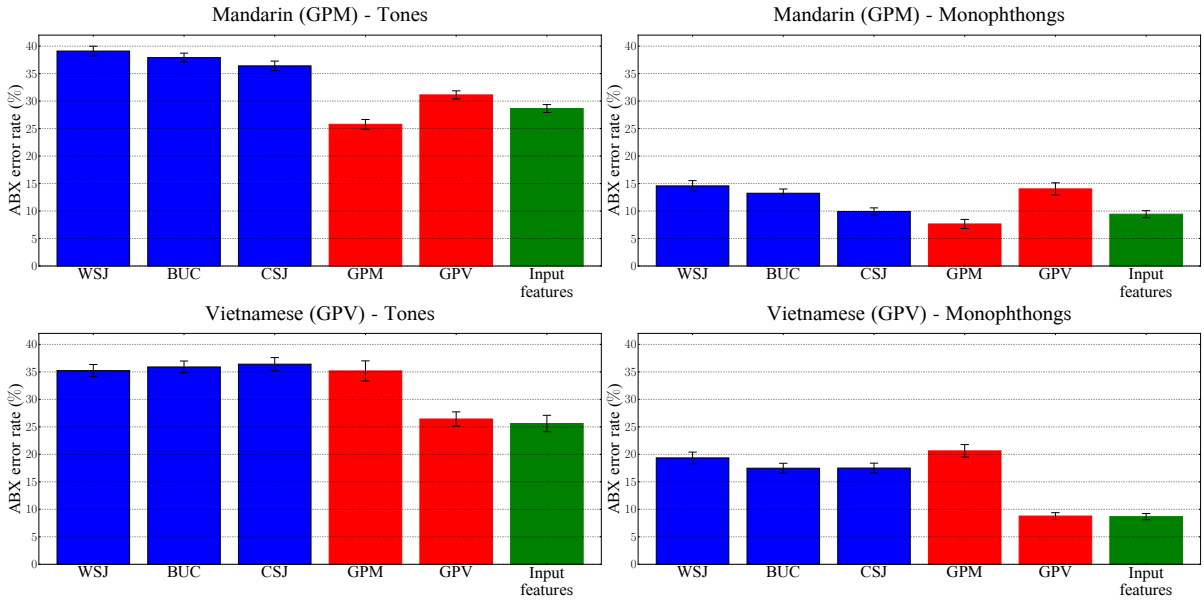


Figure 4.4: Average ABX error-rates obtained on the GPM corpus (top) and on the GPV corpus (bottom) for tone discrimination (left) and for monophthong discrimination as a control (right). Error bars indicates mean plus and minus one standard deviation and were obtained by bootstrap resampling of the scores at the level of the speakers.

The discriminability of tones is of interest because two of the four languages tested are tonal, namely Mandarin and Vietnamese. To measure the discriminability of tones we only consid-

ered phonetic contrasts that were purely tonal, i.e. obtained from two different tones applied to the same vowel. Not all possible tone contrasts occurred with every vowel in the database. We restricted ourselves to monophthongs when computing the average discriminability for all purely tonal contrasts, because a greater variety of tone contrasts occurred with monophthongs. For Mandarin, we were able to obtain discriminability scores for all tone contrasts for every monophthong, excepted tone contrasts involving the neutral tone which only occurred with the /i/ and /ə/ vowels. There were more missing contrasts in the case of Vietnamese, in which there is both more monophthongs (11 instead of 7 for Mandarin) and more tones (6 instead of 5 for Mandarin). As a control, we also computed the discriminability of monophthongs differing only in quality but with the same underlying tone. The results are plotted in Figure 4.4. The first observation is that tone distinctions are much harder to make than vowel quality distinction (25 to 40 % error rate for tone distinctions and 5 to 20 % error rate for vowel quality distinctions). In all cases, the ASR model trained on the matching corpus is much more performant than all the others ASR models, even though the improvement over input features is small or inexistent. Cross-linguistically, the model trained on CSJ is the most efficient to discriminate both Mandarin and Vietnamese monophthongs. The GPM model is the worse model at discriminating Vietnamese monophthongs and the GPV model is the worse model - tied with the WSJ model- at discriminating Mandarin monophthongs. In contrast, the GPV model is much better than the models trained on non-tonal languages at discriminating Mandarin tones, although it is slightly worse than the input features. There is no such effect in the case of the GPM model, however, which is on par with the three models trained on non-tonal languages for Vietnamese tone discrimination. Although the idea that speaker of non-tonal languages should be worse at discriminating tones than speakers of tonal languages, even cross-linguistically, is attractive, there has been conflicting reports. In certain cases, native speakers of non-tonal languages have been shown to be as bad as native speakers of a tonal language in discriminating tones from another tonal language. For example, [225] found that native speakers of Mandarin were not better than native speakers of English at discriminating Cantonese tones although native speakers of Cantonese were better than native speakers of English at discriminating Mandarin tones. Cantonese, like Vietnamese, has a more complex tone system than Mandarin, which might explain why we find a similar pattern of results. Empirical data comparing, in particular, the

discriminability of Vietnamese tones for Mandarin and English or Japanese listeners would be needed however to confirm this possibility.

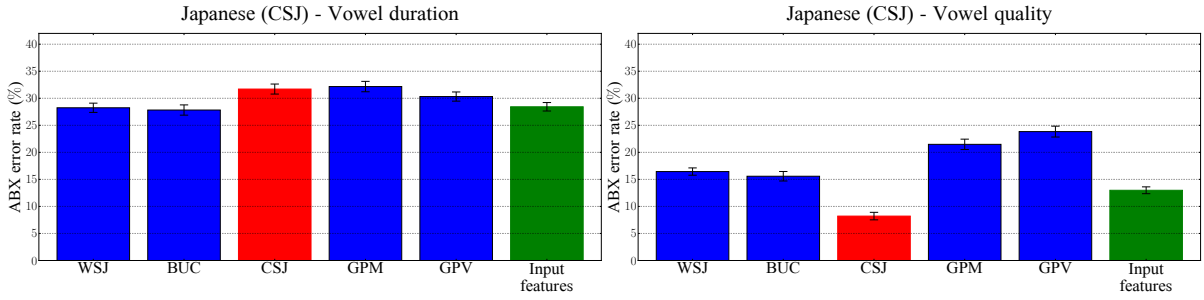


Figure 4.5: ABX error-rates obtained on the CSJ corpus for vowel duration (left) and vowel quality as a control (right). Error bars indicates mean plus and minus one standard deviation and were obtained by bootstrap resampling of the scores at the level of the speakers.

To finish our investigation of the properties of GMM-HMM ASR systems as models of cross-linguistic phonetic category perception, we consider cases showing the limits of these models. We begin with the case of vowel duration contrasts in Japanese. Each of the 5 vowels of Japanese occurs in a short and a long version, which are contrastive. In American English, certain vowels systematically differ by their length (typically tense-lax pairs, like /iː/-/ɪ/), but length differences are always coupled with differences in vowel quality and there are no pure length contrast. Native speakers of American English have been reported to have difficulty discriminating Japanese vowels which differ only in length [227, 228]. Accordingly, we expect that the model trained on the CSJ corpus should be much better than the models trained on the WSJ and BUC corpora at discriminating vowel duration contrasts. We computed the ABX error rate for each model on the CSJ corpus averaged over all vowel contrasts differing only in duration. We also computed as a control the ABX error rate for each model on the CSJ corpus averaged over all vowel contrasts differing only in quality. The results are plotted in Figure 4.5. We find a pattern of result opposite to the one expected. The model trained on CSJ appears to be the worst of all models tested at discriminating Japanese vowel duration and the two models trained on American English obtain the best results. The results on vowel quality discrimination match expectations better, with the model trained on CSJ being much better than all other models at discriminating the vowel quality of Japanese vowels. There are several potential sources for the discrepancy observed between predictions and results, including at least: the ASR model, the model of discrimination task and the speech sounds use to compute the

discrimination scores. The most likely candidate to explain the discrepancy observed between predictions and results is the GMM-HMM ASR model notorious difficulty at modeling segment durations properly. This is due to the Markov assumption of conditional independence made in Hidden Markov Models (HMM). While this assumption is the key to the efficient statistical inference procedures available for HMM, it also strongly constrains the form of the distribution of durations that can be modeled [230]. To check whether this is the right interpretation of the discrepancy between predictions and results, it would be interesting to test more recent ASR systems based on Deep Neural Networks (DNN) which do not suffer from the same limitation.

Finally, we look at four vocalic contrasts of American English (/iː/-/ɪ/, /ɛ/-/æ/, /ʌ/-/ɑː/, /æ/-/ɑː/) for which explicit ABX discrimination scores have been measured in both native speakers of Mandarin [229] and native speakers of Japanese [213] (in two separate studies). The patterns of discriminability obtained in humans are not necessarily easy to predict from the phonology of the different languages involved, but they are very different for the Japanese and Mandarin speakers. In particular the /ʌ/-/ɑː/ contrast is the hardest and the /æ/-/ɑː/ contrast is the easiest contrast for the Mandarin native speakers, while the /ʌ/-/ɑː/ contrast is relatively easy and the /æ/-/ɑː/ contrast is the hardest for the Japanese native speakers. We should observe a similar pattern for the models trained on the CSJ and GPM corpora and tested on the same contrasts of American English. More generally, the pattern of discriminability obtained with Japanese -respectively Mandarin- native speakers should be closer to the pattern of discriminability obtained with the model trained on the CSJ -respectively GPM- and tested on the WSJ or BUC corpus than to those obtained with any other model. The ABX error rates for these four contrasts estimated on both the WSJ and BUC corpus for each model along with the human results are plotted in Figure 4.6. The results do not appear to match the predictions. If anything, the patterns of discriminability obtained with the GPM model on both the WSJ and BUC corpora appear closer to the pattern obtained with Japanese native speakers and the patterns of discriminability obtained with the CSJ model on both the WSJ and BUC corpora appear closer to the pattern obtained with Mandarin native speakers. But overall, the patterns of discriminability obtained with the various ASR models as well as with the input features appear more similar to each other than to any of the patterns obtained in humans. It is not as easy to find a cause for the discrepancy observed between the predictions and the results as in

the previous case of the discrimination of vowel duration in Japanese. There are several possible sources for this discrepancy, including at least the ASR model, the model of discrimination task and the speech sounds use to compute the discrimination scores. Since some of the contrast considered involve segment with different lengths, the poor capacity of HMM-based systems to model segment durations might play a role here too, but there are many other possibilities. Other aspects of the ASR models, like the choice of using diagonal-covariance GMM for modeling the acoustic space, might play a role or the particular choice of input features, with which all the ASR models appear very correlated. The modeling of the discrimination task might also play a role. For example, it is not known whether the patterns of discriminability observed in humans remain stable if the inter-stimuli interval is varied or if participants are allowed to listen to the sounds several times for example. If it is not the case, the simple model of ABX discrimination task that we use, which is not able to account for these kinds of effects, would need to be improved. Another aspect of the modeling of the discrimination task is the choice of a notion of dissimilarity between the representations. Perhaps better results would be obtained by using a dissimilarity that only uses the category labels and the likelihood of the most likely category, as suggested by PAM, instead of using the whole posteriorgams. Or perhaps we should employ a dissimilarity that exploits even more information, such as an earth-mover's distance on the posteriorgrams. Finally, and this is probably the first thing to test, both the ASR model and the discrimination task model might be correct, but the stimuli used might be inadequate. Indeed the results in humans have been obtained with ad hoc stimuli, where all the vowels were recorded in citation form in a similar consonantic context (/h/-/d/ for the Japanese native speakers and /d/-/p/ for the Mandarin native speakers). They were also pronounced by a single speaker, a male for the study with Japanese native speakers and a female for the study with Mandarin native speakers. In comparison, we estimate discriminability of the contrasts for the different models using stimuli from a corpus of spontaneous, continuous, speech and average the results over many speakers and all phonetic contexts that we find. These differences might play an important role, since, as we mentioned previously low-level phonetic factors have been shown to affect in certain cases the patterns of discriminability of speech sounds [214]. To test this latest idea, we only need to gather the original stimuli used in the behavioral study and put them as input to our ASR models.

4.4 Conclusion

In this study, we used ABX discriminability measures to show that standard GMM-HMM ASR systems, viewed as computational models of human speech processing abilities, can account for a variety of empirically observed effects in cross-linguistic phonetic category perception by monolingual speakers. We showed that these systems can account for two types of *global* effects: first, that the phonetic categories of a language are globally harder to discriminate for non-native speakers than for native speakers and second, that the pattern of confusions between phonetic categories for non-native speakers is specific to their native language (e.g. native speakers of Japanese do not make the same confusions between phonetic categories of American English than native speakers of French). We also showed that GMM-HMM ASR systems can account for two specific *local* effects: the high confusability of American English /r/ and /l/ for native speakers of Japanese and the increased cross-linguistic confusability of stop contrasts based on VOT when compared to stop contrasts based on place for the Japanese-Mandarin language pair. We also established that GMM-HMM ASR systems predict that native speakers of Vietnamese are better than native speakers of American English or Japanese in discriminating the tones of Mandarin, but that native speakers of Mandarin are neither better nor worse than native speakers of American English or Japanese in discriminating the tones of Vietnamese. Finally, we considered two local effects that were not correctly predicted by GMM-HMM ASR systems: the higher confusability for native speakers of American English than for native speakers of Japanese of Japanese vowels differing only in duration and the differences in the pattern of discriminability of four vocalic contrasts of American English for native speakers of Japanese and Mandarin.

The approach we followed is innovative in at least two ways. First, we really model how the perceptual effects observed arise from an adaptation of speech processing abilities to the native language. This is in contrast to classical models of cross-linguistic phonetic category perception [202–209], which rely on externally provided information regarding how foreign phonetic categories map onto native ones in order to make predictions. While ASR technology and more generally machine learning tools have been available for some time and are developing very fast, they had not received a lot of attention as potential models of cross-linguistic phonetic category

perception in monolingual speakers. The two existing studies that we found [214, 215], only modeled isolated phones or syllables with limited phonetic variability, whereas we use natural continuous speech, and each of the studies only looked at one L1/L2 pair, whereas we considered eight such pairs. Considering more L1/L2 pairs, in particular, improved greatly the variety of the effects that could be investigated, and allowed to perform more controls, greatly improving the interpretability and strength of the results. Second, we evaluate models with ABX discriminability measures. This is more general than the classical model of discrimination tasks described by PAM [202–204] and relevant for discrimination results obtained both in adults and in infants or animals (see Sections 2.3.1 and 2.3.2). ABX discriminability measures can be computed for representations of speech in any format, provided a measure of dissimilarity can be defined. In comparison, PAM can only be applied to representations in terms of category labels and category goodness. This flexibility of ABX discriminability measures was already useful in this study, as it allowed us to use the MFCC and pitch features used as input to the ASR systems as a control. More generally this flexibility is highly desirable because the nature of the mental representations of speech signal in humans remains controversial. It is therefore very useful to be able to evaluate and compare the plausibility of different hypotheses in a common framework.

There are many avenues for future work. The ability of the models to account for more perceptual effects can be tested, using the exact same ABX task or introducing other relevant tasks. For example, a task comparing CC consonant clusters of American English with CVC triphones of the same language can be used to test for potential effects of epenthesis in the model trained on the CSJ. It is also straightforward to apply our methodology to other large corpora of recorded speech in the same or other languages. Beyond testing for known effects, exploratory analyses could also be used to suggest new behavioral studies from the models. For example, it would be interesting to test empirically whether the prediction that native speakers of Mandarin have no benefit in discriminating the tones of Vietnamese over native speakers of English and Japanese. It would also be interesting to try and match as closely as possible the experimental conditions in which specific effects have been observed. For example, the difficulty of Japanese native speakers with the /r/-/l/ contrast of American English is known to be more pronounced when these segments are in syllable-initial position than in syllable-final position. This could

be taken into account by introducing the position in syllable as a BY factor in the design of the ABX task. Going further, the exact same stimuli as those used in the behavioral experiments that established the perceptual effects of interest could be used to probe the models. Also, it could be interesting to model discrimination tasks in more detail (cf. Section 2.3.1). Different models, different format of speech representation, different notions of dissimilarity between these representations could also be tested. For example, DNN-based models, which should not suffer from the limitations of HMM-based models for duration contrasts, might provide more accurate models. Testing DNN models would also be interesting from the point of view of ASR. Indeed, while the limits of HMM-GMM systems, e.g. for modeling duration, are well-known by now, the strengths or weaknesses of the more recent DNN systems have been less studied. Finally, because they are trained in a supervised fashion, ASR systems have the potential to model phonetic category *perception* in adults, but not phonetic category *acquisition* in infants. It would thus be interesting to test more plausible models trained in an unsupervised or weakly supervised fashion.

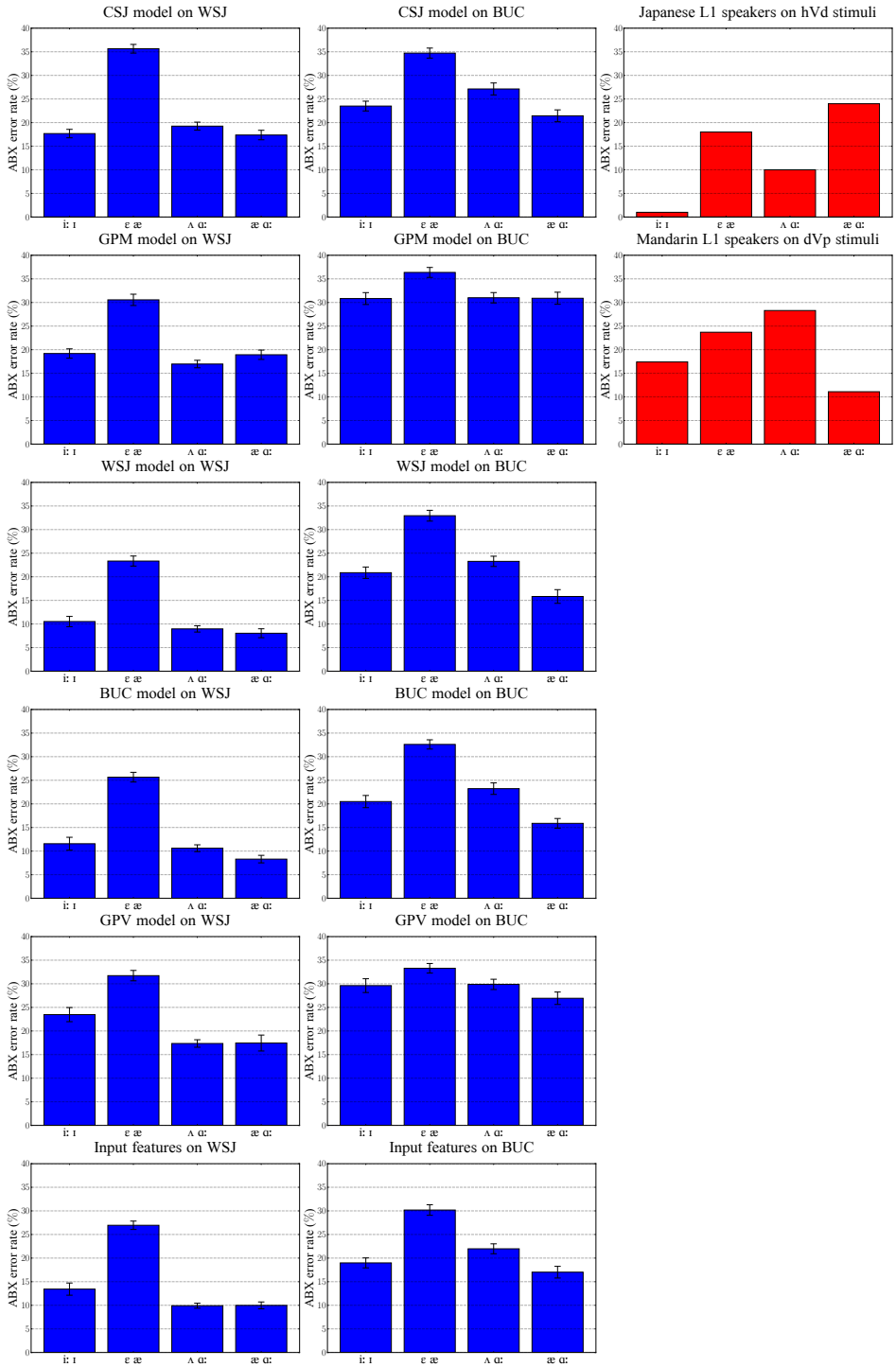


Figure 4.6: ABX error-rates obtained with ASR models on the WSJ corpus (left) and on the BUC corpus (middle) and with humans using ad hoc stimuli (right). The top row contains results for models trained on Japanese and for native speakers of Japanese. The second row contains results for models trained on Mandarin and for native speakers of Mandarin. The remaining rows contain the results for the other ASR models and the input features as controls. No human data is available for these controls. Error bars indicates mean plus and minus one standard deviation and were obtained by bootstrap resampling of the scores at the level of the speakers.

Conclusion

There were two main scientific contributions in this thesis. The first contribution was to introduce ABX-discriminability measures as a versatile tool for studying representations of categorical data. We provided the foundations for a principled usage of these measures in Chapter 1, through the study of their formal, statistical and computational properties and the comparison of these properties with those of existing alternatives. We then illustrated the practical interest of the measures by introducing three broad families of applications in Chapter 2: characterizing the performance of a system operating with limited explicit supervision in representing a category structure of interest; modeling human or animal behavior in a discrimination task; providing rich descriptive measurements for datasets annotated with category labels. We provided specific examples in the particular case where the data is constituted of large collections of speech recordings and category labels are obtained from transcriptions of these recordings.

The second contribution was to apply ABX-discriminability measures to the evaluation of models of human phonetic category processing at birth and in adulthood. In Chapter 3, faced with the scarcity of truly constraining direct empirical evidence regarding speech perception at birth, we proposed to design models based on converging evidence from auditory psychology and physiology, speech engineering and phonetic sciences and to use ABX-discriminability measures to evaluate them in terms of their functional ability to represent phonetic categories. We implemented this approach in the particular case of a family of speech features extraction methods initially developed in the context of speech engineering. In Chapter 4, we evaluated the potential of classic HMM-GMM Automatic Speech Recognition (ASR) systems as models of phonetic category processing in adults. Using 5 corpora of speech recordings in 4 different languages, we applied ABX-discriminability measures to show that these models predict, at least qualitatively, a number of effects in cross-linguistic phonetic category perception. They predict both global and local effects. We discussed two main global effects: first, phonetic contrasts are on average easier to discriminate in the training language of the system (its *mother tongue*) than in the other languages; second, the observed pattern of confusion between phonetic contrasts is largely determined by the training language/testing language pair. Local effects include the increased confusability of American English /r/-/l/ for models trained on Japanese when compared to models trained on American English and the higher loss in discriminability for stop consonants differing in voice-onset-time than for stop consonants differing in place of articulation when a

model trained on Japanese is tested on Mandarin or vice-versa. Other local effects were not correctly predicted by the models however. In particular, we could not reproduce certain effects regarding the discriminability of phonetic contrasts based on vowel duration and/or quality.

ABX-discriminability measures have already been applied in a number of published studies [35, 38, 95–98, 101–103, 105–108, 115, 231–234] and we provided strong methodological foundations for their use in Chapter 1 and Chapter 2, but some fundamental work remains to be done. While we studied in detail the case of ABX-discriminability measures between two categories, the cases of ABX-discriminability measures between multiple categories or structured categories deserve to be more thoroughly investigated than what we had time for. In particular, obtaining optimality results and deriving confidence intervals for these cases would be useful. Studying differences between ABX-discriminability measures obtained on paired samples would also be of interest, as such differences commonly arise when contrasting the behavior of several systems operating on a same set of stimuli. Another important direction for future work is to apply ABX-discriminability measures to other signals than audio recordings of speech and/or other category structures than phones, words, speakers and languages.

Regarding our work on modeling phonetic category processing at birth, a number of control experiments remain to be performed to confirm our results. Then, of course, the process can be extended to study more detailed models of auditory processing (e.g. phenomenological models accounting for more of the observed non-linearities in hearing than just dynamic-range compression or physical models of the cochlea) or more advanced speech signal processing methods (e.g. adding deltas coefficients or using features based on the modulation spectrum), allowing to assess the functional impact on the discriminability of phonetic categories of varying degree of abstraction in the models from the point of view of a system learning without explicit supervision. Of course, another important direction for future work will be to test whether the proposed models can account for the results on categorical perception of certain phonetic contrasts at birth [127–129], which are probably the only truly constraining empirical results presently available regarding phonetic category processing at birth. Regarding our approach to the study of phonetic category processing in adults, we only scratched the surface of the results that can be obtained. There are many more documented effects in cross-linguistic phonetic category perception than those we tested. A large number of these effects involve other languages

than the one we tested, but large annotated corpora of speech recordings are available for many languages, so that our approach extends straightforwardly. More detailed modeling of the experimental results to be predicted by the models might also be interesting. For example, applying the models to the exact stimuli used in the experiments instead of using phonetically matched stimuli extracted from a different corpus would be an important control. Another important remaining question is whether more recent (and more performant) ASR systems based on artificial neural networks provide better models of phonetic category processing in humans than the GMM-HMM architectures we tested. In particular, it would be interesting to see whether they fare better at predicting effects involving segment durations. Finally, our work in Chapter 3 and Chapter 4 paves the way for a thorough study of models of phonetic category acquisition during infancy. In particular, it provides a principled way to evaluate these models in a detailed and exhaustive fashion, which contrasts with the very limited and *ad hoc* testing that has to date been conducted on proposed models [3–14].

Appendix A

Proofs of results from Chapter 1

Contents

A.1 ABX discriminability for two categories

In Section 1.1.3.1, we introduced an estimator $\hat{\theta}$ for the ABX-discriminability. Several of the proofs we provide in this section rely on the identification of this estimator as a 2-sample U-statistic. More specifically, we have:

Property 16. *$\hat{\theta}$ is a U-statistic of order 2 and degree $(2, 1)$ with symmetric kernel:*

$$\Phi_d : a, x, b \mapsto \frac{1}{2}(\phi_d(a, b, x) + \phi_d(x, b, a)).$$

See [235], Section 1.2 for definitions of multi-sample U-statistics, their order and degree. See Definition 5, for the definition of ϕ_d.

Proof. Property 16

One simply needs to rewrite $\hat{\theta}(d, \mathbf{x}, \mathbf{y})$ as:

$$\frac{1}{\binom{n}{2}\binom{m}{1}} \sum_{a,x \in \mathcal{C}(\mathbf{x})} \sum_{b \in S(\mathbf{y})} \frac{1}{2} \left(\mathbb{1}_{d(a,x)<d(b,x)} + \mathbb{1}_{d(x,a)<d(b,a)} + \frac{1}{2} \left(\mathbb{1}_{d(a,x)=d(b,x)} + \mathbb{1}_{d(x,a)=d(b,a)} \right) \right),$$

where $\mathcal{C}(x)$ is the multiset $\{x_r, x_s \mid 1 \leq r < s \leq n)\}$ and $S(y)$ is as in definition 6. □

A.1.1 Point estimation

A.1.1.1 Statistical properties

As a U-statistic, $\hat{\theta}$ is necessarily unbiased, but this result is very easy to obtain even without referring to U-statistics theory as we show below.

Proof. Property 4 (Unbiasedness)

By linearity of the expectation, we have:

$$\mathbb{E}_{\mathbf{x},\mathbf{y} \sim \mathbb{P}^{\otimes m} \otimes \mathbb{Q}^{\otimes n}} [\hat{\theta}(d, \mathbf{x}, \mathbf{y})] = \frac{1}{m(m-1)n} \sum_{a \in S(\mathbf{x})} \sum_{b \in S(\mathbf{y})} \sum_{x \in S(\mathbf{x}) \setminus \{a\}} \mathbb{E}_{a,b,x \sim \mathbb{P} \otimes \mathbb{Q} \otimes \mathbb{P}} \left[\mathbb{1}_{d(a,x)<d(b,x)} + \frac{1}{2} \mathbb{1}_{d(a,x)=d(b,x)} \right].$$

Then Property 3 yields:

$$\begin{aligned} \mathbb{E}_{\mathbf{x},\mathbf{y} \sim \mathbb{P}^{\otimes m} \otimes \mathbb{Q}^{\otimes n}} [\hat{\theta}(d, \mathbf{x}, \mathbf{y})] &= \frac{1}{m(m-1)n} \sum_{a \in S(\mathbf{x})} \sum_{b \in S(\mathbf{y})} \sum_{x \in S(\mathbf{x}) \setminus \{a\}} \mathcal{D}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q}) \\ &= \mathcal{D}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q}). \end{aligned}$$

□

Proof. Property 5 (Consistency)

Since $|\Phi_d| \leq 1$, Theorem 3.2.1 from [235] pp. $97 - 98$ applies. □

Proof. (Theorem 1 (Efficiency)

Let us note $\mathcal{P}$ the set of all probability measures on (E, Π) and let us define $\mathcal{S}_{m,n} = \{P^{\otimes n} \otimes Q^{\otimes m} \mid P \in \mathcal{P},\ Q \in \mathcal{P}\}$. It is a classic result that the order statistics $(S(\mathbf{x}), S(\mathbf{y}))$ are sufficient for $\mathcal{S}_{m,n}$. By combining the result from [236], that $\mathcal{P}$ is symmetrically complete of all order and

Landers and Rogge result's on the conservation of completeness through cartesian products (see [237] proposition 1.5.6. p.19), we obtain that the order statistics $(S(\mathbf{x}), S(\mathbf{y}))$ are also complete for $\mathcal{S}_{m,n}$. Since $\hat{\theta}$ is unbiased for all distributions in $\mathcal{S}_{m,n}$ and is a measurable function, invariant by permutation of $\mathbf{x}$ and $\mathbf{y}$ (i.e. whose values depend only on the order statistics $(S(\mathbf{x}), S(\mathbf{y}))$), the Rao-Blackwell-Lehmann-Scheffe theorem (see for example [237] theorem 3.2.5 p.106) applies and yields the desired result.

□

A.1.1.2 Computational properties

Proof. Property 6

For any i in $\{1, 2, ..., m\}$, let us note σ_i a permutation of $\{1, 2, ..., m-1\}$ which sorts the elements of $(x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_m)$ in order of increasing dissimilarity to x_i according to dissimilarity function d.

In the absence of ties, we have:

$$\begin{aligned}
\hat{\theta}(d, \mathbf{x}, \mathbf{y}) &= \frac{1}{m(m-1)n} \sum_{a \in S(\mathbf{x})} \sum_{b \in S(\mathbf{y})} \sum_{x \in S(\mathbf{x}) \setminus \{a\}} \mathbb{1}_{d(a,x) < d(b,x)} \\
&= \frac{1}{m(m-1)n} \sum_{x \in S(\mathbf{x})} \sum_{a \in S(\mathbf{x}) \setminus \{x\}} \sum_{b \in S(\mathbf{y})} \mathbb{1}_{d(a,x) < d(b,x)} \\
&= \frac{1}{m(m-1)n} \sum_{i=1}^{m} \sum_{a \in S(\mathbf{x}) \setminus \{x\}} \sum_{b \in S(\mathbf{y})} \mathbb{1}_{d(a,x_i) < d(b,x_i)} \\
&= \frac{1}{m(m-1)n} \sum_{i=1}^{m} \sum_{a \in S(\mathbf{x}) \setminus \{x\}} \#\{b \in S(\mathbf{y}) \mid d(a, x_i) < d(b, x_i)\} \\
&= \frac{1}{m(m-1)n} \sum_{i=1}^{m} \sum_{j=1}^{m-1} \#\{b \in S(\mathbf{y}) \mid d(x_{\sigma_i(j)}, x_i) < d(b, x_i)\} \\
&= \frac{1}{m(m-1)n} \sum_{i=1}^{m} \sum_{j=1}^{m-1} (n - r_i(j) + j)
\end{aligned}$$

□

Proof. Property 7 (Computational complexity)

The dissimilarity function must be evaluated for each of the $m(m-1)$ pairs of two distinct elements from $\mathbf{x}$ and each of the mn pairs with a first element in $\mathbf{x}$ and a second element in $\mathbf{y}$. Once the dissimilarities are available, the most costly operation required for computing the

empirical ABX-discriminability according to the formula of Property 6 consists in sorting the elements of the sets $\{d(e, x_i)|\ e \in (S(\mathbf{x}) \cup S(\mathbf{y})) \setminus \{x_i\}\}$ for i in $\{1, 2, ..., m\}$. Each of these m sets is of size $m + n - 1$, so that the sorting of each set can be done in $O((m + n) \log (m + n))$ elementary operations using, for example, the fusion sort. □

A.1.1.3 Trading-off statistical efficiency for computational efficiency through a sampled estimator

Proof. Property 8

Let us consider a tuple $\mathbf{t}$ formed of the admissible triplets enumerated in an arbitrary order (the admissible triplets being the element of the multiset $\{(a, b, x) \in S(\mathbf{x}) \times S(\mathbf{y}) \times S(\mathbf{x}) \mid a \neq x\}$). Let us note N the size of $\mathbf{t}$ ($N = m(m-1)n$). We note t_i the i-th component of t. Then, we can rewrite the sampled estimator as:

$$\hat{\theta}_B(d, \mathbf{x}, \mathbf{y}) = \frac{1}{B} \sum_{i=1}^{N} w_i \phi_d(t_i),$$

where $\mathbf{w} = (w_1, w_2, ..., w_N)$ is a sample from the multinomial distribution with parameters B and $\left(\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N}\right)$.

Using the same notations for the underlying random variables as for the samples, we then have:

$$\mathrm{Var}(\hat{\theta}_B) = \frac{1}{B^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{Cov}(w_i \phi_d(t_i), w_j \phi_d(t_j)).$$

For any i in $\{1, 2, ..., N\}$:

$$\begin{aligned}
\mathrm{Cov}(w_i \phi_d(t_i), w_i \phi_d(t_i)) &= \mathrm{Var}(w_i \phi_d(t_i)) \\
&= \mathrm{Var}(w_i)\mathrm{Var}(\phi_d(t_i)) + \mathrm{Var}(w_i)\mathbb{E}(\phi_d(t_i))^2 + \mathbb{E}(w_i)^2 \mathrm{Var}(\phi_d(t_i)) \\
&= \frac{B(N-1)}{N^2}\mathrm{Var}(\phi_d(t_i)) + \frac{B(N-1)}{N^2}\mathbb{E}(\phi_d(t_i))^2 + \frac{B^2}{N^2}\mathrm{Var}(\phi_d(t_i))
\end{aligned}$$

For any i, j in $\{1, 2, ..., N\}^2$, $i \neq j$:

$$\begin{aligned}
\mathrm{Cov}(w_i\phi_d(t_i), w_j\phi_d(t_j)) &= \mathbb{E}(w_i\phi_d(t_i)w_j\phi_d(t_j)) - \mathbb{E}(w_i\phi_d(t_i))\mathbb{E}(w_j\phi_d(t_j)) \\
&= \mathbb{E}(w_iw_j)\mathbb{E}(\phi_d(t_i)\phi_d(t_j)) - \mathbb{E}(w_i)\mathbb{E}(\phi_d(t_i))\mathbb{E}(w_j)\mathbb{E}(\phi_d(t_j)) \\
&= (\mathrm{Cov}(w_i, w_j) + \mathbb{E}(w_i)\mathbb{E}(w_j))\mathbb{E}(\phi_d(t_i)\phi_d(t_j)) - \frac{B^2}{N^2}\mathbb{E}(\phi_d(t_i))\mathbb{E}(\phi_d(t_j)) \\
&= (-\frac{B}{N^2} + \frac{B^2}{N^2})\mathbb{E}(\phi_d(t_i)\phi_d(t_j)) - \frac{B^2}{N^2}\mathbb{E}(\phi_d(t_i))\mathbb{E}(\phi_d(t_j)) \\
&= \frac{B(B-1)}{N^2}(\mathrm{Cov}(\phi_d(t_i), \phi_d(t_j)) + \mathbb{E}(\phi_d(t_i))\mathbb{E}(\phi_d(t_j))) - \frac{B^2}{N^2}\mathbb{E}(\phi_d(t_i))\mathbb{E}(\phi_d(t_j)) \\
&= \frac{B(B-1)}{N^2}\mathrm{Cov}(\phi_d(t_i), \phi_d(t_j)) - \frac{B}{N^2}\mathbb{E}(\phi_d(t_i))\mathbb{E}(\phi_d(t_j))
\end{aligned}$$

Furthermore, for any i in $\{1, 2, ..., N\}$, $\mathbb{E}(\phi_d(t_i)) = \theta$ (Property 3).

Putting these three results together, we obtain:

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_B) = \frac{B(B-1)}{B^2N^2}\sum_{i=1}^{N}\sum_{j\in\{1,2,...,N\}\setminus\{i\}}\mathrm{Cov}(\phi_d(t_i), \phi_d(t_j)) - \frac{B(N-1)}{B^2N}\theta^2 \\
+ \frac{B^2 + BN - B}{B^2N^2}\sum_{i=1}^{N}\mathrm{Var}(\phi_d(t_i)) + \frac{B(N-1)}{B^2N}\theta^2.
\end{aligned}$$

Which simplifies into:

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_B) &= \frac{B-1}{BN^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{Cov}(\phi_d(t_i), \phi_d(t_j)) + \frac{1}{BN}\sum_{i=1}^{N}\mathrm{Var}(\phi_d(t_i)) \\
&= \frac{B-1}{B}\mathrm{Var}(\hat{\theta}) + \frac{1}{BN}\sum_{i=1}^{N}\mathrm{Var}(\phi_d(t_i))
\end{aligned}$$

Let us consider i in $\{1, 2, ..., N\}$. We have $t_i = (a, b, x)$ for some mutually independent random variables $a, b, x \sim \mathbb{P} \otimes \mathbb{Q} \otimes \mathbb{P}$. Then:

$$\begin{aligned}
\mathrm{Var}(\phi_d(t_i)) &= \mathbb{E}[\phi_d(t_i)^2] - \theta^2 \\
&= \mathbb{E}\left[(\mathbb{1}_{d(a,x)<d(b,x)})^2 + \left(\frac{1}{2}\mathbb{1}_{d(a,x)=d(b,x)}\right)^2 + \mathbb{1}_{d(a,x)<d(b,x)}\mathbb{1}_{d(a,x)=d(b,x)}\right] - \theta^2
\end{aligned}$$

Since $(\mathbb{1}_{d(a,x)<d(b,x)})^2 = \mathbb{1}_{d(a,x)<d(b,x)}$, $(\mathbb{1}_{d(a,x)=d(b,x)})^2 = \mathbb{1}_{d(a,x)=d(b,x)}$ and $\mathbb{1}_{d(a,x)<d(b,x)}\mathbb{1}_{d(a,x)=d(b,x)} = 0$, we obtain:

$$\begin{aligned}\mathrm{Var}(\phi_d(t_i)) &= \mathbb{E}\left(\mathbb{1}_{d(a,x)<d(b,x)} + \frac{1}{4}\mathbb{1}_{d(a,x)=d(b,x)}\right) - \theta^2 \\ &= \mathbb{E}(\phi_d(t_i) - \frac{1}{4}\mathbb{E}\left(\mathbb{1}_{d(a,x)=d(b,x)}\right) + \theta^2 \\ &= \theta(1-\theta) - \frac{1}{4}\mathbb{E}\left(\mathbb{1}_{d(a,x)=d(b,x)}\right) \\ &\leq \theta(1-\theta)\end{aligned}$$

In the end we obtain:

$$\mathrm{Var}(\hat{\theta}_B) \leq \frac{B-1}{B}\mathrm{Var}(\hat{\theta}) + \frac{\theta(1-\theta)}{B}.$$

Furthermore, since for any $x \in [0,1]$, $x(1-x) \leq \frac{1}{4}$, we deduce:

$$\mathrm{Var}(\hat{\theta}_B) \leq \frac{B-1}{B}\mathrm{Var}(\hat{\theta}) + \frac{1}{4B}.$$

□

Property 9 is trivial and we do not provide an explicit proof for it.

A.1.2 Interval estimation

Proof. Theorem 2 (Non-asymptotic confidence interval)

Because we have $0 \leq \Phi_d \leq 1$, the confidence interval can be obtained directly from the concentration inequality introduced by Hoeffding for two-sample U-statistics in Section 4*b*. of [238]. □

Proof. Theorem 3 (Bootstrap asymptotic confidence interval)

Let us note $\hat{\theta}_{m,n} := \hat{\theta}(d, \mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ is a random variable with law $\mathbb{P}^{\otimes m}$ and $\mathbf{Y}$ is a random variable with law $\mathbb{Q}^{\otimes n}$. Let us also note $\hat{\theta}^*_{m,n} := \hat{\theta}(d, \mathbf{X}^*, \mathbf{Y}^*)$, where $\mathbf{X}^*$ and $\mathbf{Y}^*$ are as in the statement of the theorem.

$\hat{\theta}$ is a U-statistic (see Property 16) and under the assumptions $\rho_{10} > 0$ and $\rho_{01} > 0$ of our theorem it is non-degenerate. Thus from Theorem 2.4 and subsequent remarks of [239], we obtain that $\frac{\hat{\theta}_{m,n}-\theta}{\sqrt{\min m,n}}$ and $\frac{\hat{\theta}^*_{m,n}-\hat{\theta}_{m,n}}{\sqrt{\min m,n}}$ converge weakly toward the same limit, a Gaussian random

variable. Since Gaussian distribution functions are continuous, Lemma 23.3 p.329 of [240] applies and yields the desired result. □

A.2 Comparison with other measures of the separation of two categories

Proof. Property 10 Let us consider a, b, x, y, four independent random variables distributed according to $\mathbb{P}$, $\mathbb{Q}$, $\mathbb{P}$ and $\mathbb{P}$ respectively. Let us note $d_1 = d(a,x)$ and $d_2 = d(b,y)$, which are independent real random variables. By definition, $F : \tau \mapsto \mathcal{F}_{\text{AX}}(d,\mathbb{P},\tau)$ is equal to 1 minus the cumulative distribution function of d_1 plus $\tau \mapsto \frac{1}{2}\mathbb{E}[\mathbb{1}_{d_1=\tau}]$. In the same way $T : \tau \mapsto \mathcal{T}_{\text{AX}}(d,\mathbb{P},\mathbb{Q},\tau)$ is equal to 1 minus the cumulative distribution function of d_2 plus $\tau \mapsto \frac{1}{2}\mathbb{E}[\mathbb{1}_{d_2=\tau}]$.

We have defined $\mathcal{D}_{\text{AX}}$ under the assumptions that T is $\mathcal{C}^1$ and F is a $\mathcal{C}^1$-diffeomorphism. It is easy to show that the continuity of these functions in particular implies that for all $\tau \in \mathbb{R}$, $\frac{1}{2}\mathbb{E}[\mathbb{1}_{d_1=\tau}] = 0$ and $\frac{1}{2}\mathbb{E}[\mathbb{1}_{d_2=\tau}] = 0$. From this we can deduce that the cumulative distribution functions of d_1 and d_2 are absolutely continuous with respect to the measure of Lebesgue and that $-F'$ is a probability density for d_1 and $-T'$ is a probability density for d_2. Then we can write:

$$
\begin{aligned}
\mathcal{D}_{\text{ABXY}}(d,\mathbb{P},\mathbb{Q}) &= \int_{u\in\mathbb{R}}\int_{v\in\mathbb{R}} \mathbb{1}_{u<v}(-F'(u))(-T'(v))\lambda(du)\lambda(dv) \\
&= \int_{u\in\mathbb{R}}\int_{v\in]u,+\infty[} F'(u)T'(v)\lambda(du)\lambda(dv) \\
&= \int_{u\in\mathbb{R}}\left(\int_{v\in]u,+\infty[} T'(v)\lambda(dv)\right)F'(u)\lambda(du) \\
&= \int_{u\in\mathbb{R}}\left(-\int_{+\infty}^{u} T'(v)dv\right)F'(u)\lambda(du) \\
&= \int_{u\in\mathbb{R}} -(T(u)-0)F'(u)\lambda(du) \\
&= -\int_{u\in\mathbb{R}} T(u)F'(u)\lambda(du) \\
&= \int_{+\infty}^{-\infty} T(u)F'(u)du \\
&= \mathcal{D}_{\text{AX}}(d,\mathbb{P},\mathbb{Q})
\end{aligned}
$$

□

Proof. Theorem 4

Let us note:

$$\Delta := \mathcal{D}_{\text{ABX}}(d, \mathbb{P}, \mathbb{Q}) - \mathcal{D}_{\text{AX}}(d, \mathbb{P}, \mathbb{Q})$$
$$\phi_d(a, b, x) := \mathbb{1}_{d(a,x)<d(b,x)} + \frac{1}{2}\mathbb{1}_{d(a,x)=d(b,x)}$$
$$\psi_d(a, b, x, y) := \mathbb{1}_{d(a,x)<d(b,y)} + \frac{1}{2}\mathbb{1}_{d(a,x)=d(b,y)}$$
$$\alpha_d(a, b, x, y) := \mathbb{1}_{d(b,y)\leq d(a,x)<d(b,x)} + \frac{1}{2}\mathbb{1}_{d(b,y)\neq d(a,x)=d(b,x)}$$
$$\beta_d(a, b, x, y) := \mathbb{1}_{d(b,x)\leq d(a,x)<d(b,y)} + \frac{1}{2}\mathbb{1}_{d(b,x)\neq d(a,x)=d(b,y)}$$

Then:

$$\begin{aligned}
\Delta &= \int_{a,b,x\in E^3} \phi_d(a, b, x)\mathbb{P}(da)\mathbb{Q}(db)\mathbb{P}(dx) - \int_{a,b,x,y\in E^4} \psi_d(a, b, x, y)\mathbb{P}(da)\mathbb{Q}(db)\mathbb{P}(dx)\mathbb{P}(dy) \\
&= \int_{a,b,x\in E^3} \left(\phi_d(a, b, x) - \int_{y\in E} \psi_d(a, b, x, y)\mathbb{P}(dy)\right) \mathbb{P}(da)\mathbb{Q}(db)\mathbb{P}(dx) \\
&= \int_{a,b,x\in E^3} \int_{y\in E} \left(\phi_d(a, b, x) - \psi_d(a, b, x, y)\right) \mathbb{P}(dy)\mathbb{P}(da)\mathbb{Q}(db)\mathbb{P}(dx) \\
&= \int_{a,b,x\in E^3} \int_{y\in E} \left(\alpha_d(a, b, x, y) - \beta_d(a, b, x, y)\right) \mathbb{P}(dy)\mathbb{P}(da)\mathbb{Q}(db)\mathbb{P}(dx) \\
&= \int_{a,b,x,y\in E^4} \alpha_d(a, b, x, y)\mathbb{P}(dy)\mathbb{P}(da)\mathbb{Q}(db)\mathbb{P}(dx) - \int_{a,b,x,y\in E^4} \beta_d(a, b, x, y)\mathbb{P}(dy)\mathbb{P}(da)\mathbb{Q}(db)\mathbb{P}(dx) \\
&= p_1(d, \mathbb{P}, \mathbb{Q}) - p_2(d, \mathbb{P}, \mathbb{Q})
\end{aligned}$$

□

Proof. Theorem 5

We show the contrapositive of the theorem. If the classification accuracy is strictly less than one, then there is a point X in the test set whose closest neighbor B in the training set is not of the same class. Let us consider a point A from the training set of the same class as X (we assumed that there was at least a point of each class in the training set). Then, the triplet A, B, X is one of the triplets considered in the computation of the ABX-discriminability $\hat{\theta}$. Since $d(A, X) > d(B, X)$, we can conclude that the ABX-discriminability is strictly less than one. □

In the next proof, we assume that d is symmetric, as asymmetric dissimilarity functions do not make sense for clustering. We also suppose that a sample $\mathbf{z} = (z_1, z_2, ..., z_n) \in E^n$ from two categories as specified by category labels $\mathbf{c} = (c_1, c_2, ...c_n) \in \{1, 2\}^n$ is given.

Proof. Theorem 6 The algorithm stops once there is only two remaining clusters. Let us suppose that one of the two resulting clusters, contains a point x from $\mathbf{x}$ and a point y from $\mathbf{y}$. This implies that there exists $i \in \mathbb{N}^*$, such that at iteration i of the merging step of the single linkage procedure, there were strictly more than two clusters and the two closest points belonging to different clusters were a point x_0 from $\mathbf{x}$ and a point y_0 from $\mathbf{y}$. We have:

$$d(x_0, y_0) \geq \min_{b \in \mathbf{y}} d(x_0, b)$$

and since $\mathbf{y}$ is finite, there exists $b^* \in \mathbf{y}$, such that:

$$d(x_0, b^*) = \min_{b \in \mathbf{y}} d(x_0, b).$$

Moreover, $\hat{\theta}(d, \mathbf{x}, \mathbf{y}) = 1$ implies for all $a \in \mathbf{x} \setminus \{x_0\}$:

$$d(x_0, b^*) > d(x_0, a).$$

In the end, we obtain:

$$d(x_0, y_0) > \max_{a \in \mathbf{x} \setminus \{x_0\}} d(x_0, a).$$

Since, at iteration i the two closest points belonging to different clusters were x_0 and y_0, this implies that the cluster containing x_0 at this point in the execution of the algorithm also contains all the other elements of $\mathbf{x}$. By symmetry we also obtain that the cluster containing y_0 at iteration i also contains all the other elements of $\mathbf{y}$. This is in contradiction with the assertion that there were strictly more than two clusters at iteration i. We can therefore conclude *ad absurdio*, that the two final clusters are $\mathbf{x}$ and $\mathbf{y}$. □

Property 17 (Asymptotic solution for 2-means clustering in the line example)**.** *The asymptotic solution for* 2*-means clustering in the line example of Section 1.2.3.4 is formed of two clusters with centers:*

$$(c_1, c_2) = \begin{cases} \left(-\frac{2+2d+d^2}{4}, \frac{2+2d+d^2}{4}\right) & \text{if } -1 \leq d < 0 \\ \left(-\frac{d+1}{2}, \frac{d+1}{2}\right) & \text{if } d \geq 0 \end{cases}.$$

Proof. We give the proof only for the most difficult case $-1 < d < 0$, the proof for other cases

follows the same lines.

Given $d \in \mathbb{R}$, by definition 2-means solutions $(c_1, c_2) \in \mathbb{R}^2$, where we suppose $c1 \leq c2$ without loss of generality, should minimize:

$$\begin{aligned} f_d(c_1,c_2) &:= \int_{\min\left(-\frac{d}{2}-1,\frac{c_1+c_2}{2}\right)}^{\min\left(-\frac{d}{2},\frac{c_1+c_2}{2}\right)} (x-c_1)^2 dx \\ &+ \int_{\min\left(\frac{d}{2},\frac{c_1+c_2}{2}\right)}^{\min\left(\frac{d}{2}+1,\frac{c_1+c_2}{2}\right)} (x-c_1)^2 dx \\ &+ \int_{\max\left(\frac{d}{2},\frac{c_1+c_2}{2}\right)}^{\max\left(\frac{d}{2}+1,\frac{c_1+c_2}{2}\right)} (x-c_2)^2 dx \\ &+ \int_{\max\left(-\frac{d}{2}-1,\frac{c_1+c_2}{2}\right)}^{\max\left(-\frac{d}{2},\frac{c_1+c_2}{2}\right)} (x-c_2)^2 dx. \end{aligned}$$

Let us note $D_0 = \{(c_1, c_2) \in \mathbb{R}^2 \mid c_1 \leq c_2\}$ the definition set of f_d. If $-1 < d < 0$, we can distinguish five cases:

1. $\frac{c_1+c_2}{2} < -\frac{d}{2} - 1$. Then:

$$f_d(c_1,c_2) = \int_{\frac{d}{2}}^{\frac{d}{2}+1} (x-c_2)^2 dx + \int_{-\frac{d}{2}-1}^{-\frac{d}{2}} (x-c_2)^2 dx = \frac{2}{3} + d + \frac{d^2}{2} + 2c_2^2.$$

   It follows that f_d reaches a minimum of $\frac{2}{3} + d + \frac{d^2}{2}$ on the subdomain D of D_0 defined by $\frac{c_1+c_2}{2} < -\frac{d}{2} - 1$ (in any point (c_1, c_2) with $c_2 = 0$, $c_1 < -d - 2$).

2. $-\frac{d}{2} - 1 \leq \frac{c_1+c_2}{2} \leq \frac{d}{2}$. Then:

$$\begin{aligned} f_d(c_1,c_2) &= \int_{-\frac{d}{2}-1}^{\frac{c_1+c_2}{2}} (x-c_1)^2 dx + \int_{\frac{d}{2}}^{\frac{d}{2}+1} (x-c_2)^2 dx + \int_{\frac{c_1+c_2}{2}}^{-\frac{d}{2}} (x-c_2)^2 dx \\ &= \frac{2}{3} + d + \frac{d^2}{2} + c_1^2 + c_2^2 + (c_1-c_2)\left[1 + \left(\frac{c_1+c_2}{2}\right)^2 + d\left(1 + \frac{c_1+c_2}{2}\right) + \frac{d^2}{4}\right]. \end{aligned}$$

   Thus f_d is polynomial and therefore $\mathcal{C}^1$ on the subdomain D of D_0 defined by $-\frac{d}{2} - 1 \leq \frac{c_1+c_2}{2} \leq \frac{d}{2}$. This subdomain is closed and it is easy to show that the gradient of f_d on the interior of D has no zeroes, so that the extrema of the restriction of f_d to D are necessarily at the boundary of D. The boundary is the union of $B_1 = \{(c_1, c_2) \in D_0 \mid c_2 = -d-2-c_1\}$ and $B_2 = \{(c_1, c_2) \in D_0 \mid c_2 = d - c_1\}$. It is easily shown that f_d reaches a minimum of $\frac{2}{3} + d + \frac{d^2}{2}$ on B_1 in $(-d-2, 0)$ and a minimum of $\frac{1}{6} - \frac{d^2}{2} - d^3 - \frac{d^4}{2}$ on B_2 in $\left(-\frac{1+d^2}{2}, \frac{(1+d)^2}{2}\right)$. Studying the sign of the difference of these quantities for $-1 < d < 0$, we obtain that f_d

reaches a global minimum of $\frac{1}{6} - \frac{d^2}{2} - d^3 - \frac{d^4}{2}$ on D in $\left(-\frac{1+d^2}{2}, \frac{(1+d)^2}{2}\right)$.

3. $\frac{d}{2} \leq \frac{c_1+c_2}{2} \leq -\frac{d}{2}$. Then:

$$\begin{aligned} f_d(c_1,c_2) &= \int_{-\frac{d}{2}-1}^{\frac{c_1+c_2}{2}} (x-c_1)^2 dx + \int_{\frac{d}{2}}^{\frac{c_1+c_2}{2}} (x-c_1)^2 dx + \int_{\frac{c_1+c_2}{2}}^{\frac{d}{2}+1} (x-c_2)^2 dx + \int_{\frac{c_1+c_2}{2}}^{-\frac{d}{2}} (x-c_2)^2 dx \\ &= \frac{2}{3} + d + \frac{d^2}{2} + c_1^2 + c_2^2 + (c_1-c_2)\left[1 + \frac{(c_1+c_2)^2}{2} + d + \frac{d^2}{2}\right]. \end{aligned}$$

Thus f_d is polynomial and therefore $\mathcal{C}^1$ on the subdomain D of D_0 defined by $\frac{d}{2} < \frac{c_1+c_2}{2} < -\frac{d}{2}$. The gradient of f_d on the interior of D reaches 0 only in $p = \left(-\frac{2+2d+d^2}{4}, \frac{2+2d+d^2}{4}\right)$. Since D is closed, the extrema of the restriction of f_d to D are necessarily at the boundary of D or in p. The boundary is the union of $B_1 = \{(c_1,c_2) \in D_0 \mid c_2 = d - c_1\}$ and $B_2 = \{(c_1,c_2) \in D_0 \mid c_2 = -d - c_1\}$. We already established that f_d reaches a minimum of $\frac{1}{6} - \frac{d^2}{2} - d^3 - \frac{d^4}{2}$ on B_1 in $\left(-\frac{1+d^2}{2}, \frac{(1+d)^2}{2}\right)$. On B_2 it can be shown to reach the same minimum in $\left(-\frac{(1+d)^2}{2}, \frac{1+d^2}{2}\right)$. In p the value of f_d is $\frac{1}{6} - \frac{d^2}{2} - \frac{d^3}{2} - \frac{d^4}{8}$, which is always strictly lower than $\frac{1}{6} - \frac{d^2}{2} - d^3 - \frac{d^4}{2}$ for $-1 < d < 0$. Therefore, f_d reaches a global minimum of $\frac{1}{6} - \frac{d^2}{2} - \frac{d^3}{2} - \frac{d^4}{8}$ on D in $\left(-\frac{2+2d+d^2}{4}, \frac{2+2d+d^2}{4}\right)$.

4. $-\frac{d}{2} \leq \frac{c_1+c_2}{2} \leq \frac{d}{2} + 1$. Then:

$$\begin{aligned} f_d(c_1,c_2) &= \int_{-\frac{d}{2}-1}^{-\frac{d}{2}} (x-c_1)^2 dx + \int_{\frac{d}{2}}^{\frac{c_1+c_2}{2}} (x-c_1)^2 dx + \int_{\frac{c_1+c_2}{2}}^{\frac{d}{2}+1} (x-c_2)^2 dx \\ &= \frac{2}{3} + d + \frac{d^2}{2} + c_1^2 + c_2^2 + (c_1-c_2)\left[1 + \left(\frac{c_1+c_2}{2}\right)^2 + d\left(1 - \frac{c_1+c_2}{2}\right) + \frac{d^2}{4}\right]. \end{aligned}$$

Following the same line of reasoning as in the previous cases, we obtain that f_d reaches a global minimum of $\frac{1}{6} - \frac{d^2}{2} - d^3 - \frac{d^4}{2}$ in $\left(-\frac{(1+d)^2}{2}, \frac{1+d^2}{2}\right)$ on the subdomain of D_0 defined by $-\frac{d}{2} \leq \frac{c_1+c_2}{2} \leq \frac{d}{2} + 1$.

5. $\frac{d}{2} + 1 < \frac{c_1+c_2}{2}$. Then:

$$f_d(c_1,c_2) = \int_{-\frac{d}{2}-1}^{-\frac{d}{2}} (x-c_1)^2 dx + \int_{\frac{d}{2}}^{\frac{d}{2}+1} (x-c_1)^2 dx = \frac{2}{3} + d + \frac{d^2}{2} + 2c_1^2.$$

It follows that f_d reaches a minimum of $\frac{2}{3} + d + \frac{d^2}{2}$ on the subdomain D of D_0 defined by $\frac{d}{2} + 1 < \frac{c_1+c_2}{2}$ (in any point (c_1,c_2) with $c_1 = 0$, $c_2 > d + 2$).

□

Appendix B

Short-term power spectrum and auditory models

Contents

B.1 Reinterpreting Short-Term Power Spectrum

In this section, we show that STPS extraction followed by frequency rescaling on a lin-log scale and dynamic-range compression is approximately equivalent to convolution with a real-valued filterbank followed by compression of the dynamic range and envelope extraction. Let us first show that STPS extraction can be seen as a sequence of two distinct operations: a convolution of the input signal with a real-valued filterbank followed by the extraction of a slowly-varying envelope from the output of each filter in the filterbank. More specifically, we show that under reasonable assumptions, the Short-Term Power Spectrum of a signal is constituted of the squared Hilbert envelopes of convolutions of this signal with modulated cosines.

Let us consider a sampled input signal $x(n)$, $n \in \mathbb{Z}$. We suppose that x has finite support (this assumption can be made because all sounds considered in practical applications have finite durations). We consider a time-localizing window v for the short-term Fourier Transform, also with finite support and we suppose that it has a low pass character with no or negligible energy in the frequency channels for which we want to obtain a short-term power spectrum measure.

By definition, the short-term power spectrum of x at frequency $\frac{\omega}{2\pi}$ and instant m based on

the time-localizing window v is:

$$p(m,\omega) := \left| \sum_{n\in\mathbb{Z}} x(n)v(n-m)e^{-i\omega n} \right|^2 .$$

Let us note c_ω the real part of ψ_ω, i.e. $c_\omega : n \mapsto v(-n)\cos\omega n$. Then we have the following result:

Theorem 7 (STPS intepretation)**.**

$$p(.,\omega) = \mathcal{E}(x \star c_\omega)^2,$$

where $\mathcal{E}(s)$ denotes the Hilbert envelope of signal s.

Proof. Theorem 7

We can rewrite $p(m,\omega)$ in order to see it as the squared modulus of a convolution of the input signal with a complex exponential modulated by the time-reversed version of the time-localizing window:

$$p(m,\omega) = \left| e^{-i\omega m} \sum_{n\in\mathbb{Z}} x(n)v(-(m-n))e^{i\omega(m-n)} \right|^2 = |x \star \psi_\omega(m)|^2,$$

with $\psi_\omega : m \mapsto v(-m)e^{j\omega m}$.

By definition the Hilbert envelope of s is the magnitude of the analytic signal (see for example [241], Chapter 2) associated to s, i.e.:

$$\mathcal{E}(s) = |s + is \star h|,$$

where h is the impulse response for the discrete Hilbert transform (see for example [242]). So,

$$\mathcal{E}(x \star c_\omega) = |x \star c_\omega + i(x \star c_\omega) \star h|.$$

Since both the time-localizing window and the input signal have finite support, all the sums involved in computing $(x \star c_\omega) \star h$ and $x \star (c_\omega \star h)$ are finite and it is easily shown that both quantities are equal. By linearity we then obtain:

$$\mathcal{E}(x \star c_\omega) = |x \star (c_\omega + ic_\omega \star h)|.$$

We will show next that under the assumptions we made on the time-localizing window v, we have $c_\omega \star h = s_\omega$, with $s_\omega : n \mapsto v(-n) \sin \omega n$. This will complete the proof as it is easily obtained from it that

$$\mathcal{E}(x \star c_\omega)^2 = |x \star (c_\omega + i s_\omega)|^2 = |x \star \psi_\omega|^2 = p(., \omega).$$

To see that $c_\omega \star h = s_\omega$, consider that c_ω can be seen as the product of the window function with the cosine function, so that its Fourier transform $\hat{c}_\omega$ is the convolution of the Fourier transform $\hat{v}$ of the window function and of the Fourier transform of the cosine function $\hat{C}_\omega$. The Fourier transform of the cosine function is $\hat{C}_\omega : x \mapsto \frac{1}{2}\mathbb{1}_{-\omega}(x) + \frac{1}{2}\mathbb{1}_{\omega}(x)$ and the Fourier transform of the corresponding sine function is $\hat{S}_\omega : x \mapsto -\frac{1}{2i}\mathbb{1}_{-\omega}(x) + \frac{1}{2i}\mathbb{1}_{\omega}(x)$. From this, given that we assumed that $\hat{v}$ has zero or negligible power above ω (and below $-\omega$), we can deduce that $x \mapsto -i\text{sgn}(x)\hat{c}_\omega(x) = \hat{s}_\omega$, where $\hat{s}_\omega$ is the Fourier transform of s_ω, which proves the result. □

So, STPS extraction can be seen as convolution with a filterbank followed by envelope extraction. We want to show that STPS extraction followed by frequency rescaling and dynamic-range compression is approximately equivalent to convolution with a filterbank followed by dynamic range compression and envelope extraction. The next steps to obtain this result are to show that performing the envelope extraction step before or after frequency rescaling or dynamic range compression has little impact on the final output and that the convolution with a filterbank followed by frequency rescaling is equivalent to convolution with a different filterbank. We refer to [104] Section II.B, for the demonstration that frequency rescaling approximately commutes with envelope extraction and that convolution with a filterbank followed by frequency rescaling is equivalent to convolution with a different filterbank. The last thing to justify is that cubic-root dynamic range compression approximately commutes with Hilbert envelope extraction. It is not true in general that Hilbert envelope extraction commutes with power laws, but changing the order of the two operations does not appear to make a big difference in the final result in practice. It would be nice to find reasonable conditions on the input signal for which there is a theoretical guarantee that applying dynamic range compression before envelope extraction results in an envelope whose dynamic range is compressed, but we leave that for future work.

B.2 Simple phenomenological models of the cochlea

In this appendix, we explain the rationale for simple phenomenological models of the cochlea constituted of a linear system followed by a static compressive nonlinearity.

Let us first introduce a simplified view of the cochlea, appropriate to describe its mechanical properties (see for example [243], Chapter 3). The cochlea can be seen as a solid tube containing two cavities running along its length, the *scala tympani* and the *scala vestibuli*. The top of the tube is usually referred to as the base of the cochlea and the bottom as the apex of the cochlea. The two cavities communicate with each other near the apex at a place called the *helicotrema* and are completely filled with an incompressible aqueous fluid. The walls of the cavities are rigid except for a flexible stretch of tissue separating the two cavities, the *cochlear partition*, and two flexible membranes sealing the top of the cavities: the *oval window* sealing the scala tympani and the *round window* sealing the *scala vestibuli*. The cochlear partition is a complex object, but the main thing we need to know for a high-level explanation is that it becomes gradually more flexible as it goes toward the apex and that it contains two type of neural cells: Inner Hair Cells (IHC) and Outer Hair Cells (OHC). The IHC are the primary sensory neurons of audition, they detect vibrations at different points along the cochlear partition and transduce them into graded electrical potentials that are transmitted further along the sensory pathways toward the brain. The OHC are effector neurons which react to mechanical stimulations in a way that is not yet fully understood, but results in the active amplification of small vibrations of the cochlear partition. There are around $3,000$ IHC spread regularly in a single row along the cochlear partition and around $12,000$ OHC spread regularly in three parallel rows along the cochlear partition.

We are now ready to explain how the cochlea works. Note that the cochlea is a complex and delicate mechanical system whose normal mode of operation is very hard to observe experimentally and there is still some controversy about the details of how it works (see for example [244] or [245]). We only give a high-level overview of the most commonly admitted principles. Incoming sounds are transmitted through the middle ear resulting in pulling and pushing motion on the oval window in synchrony with the pressure wave of the sound, as filtered by the middle hear. Since the cochlear compartments are filled with incompressible fluid, this results in displacement of the fluid within the cochlea, inward movement at the oval window resulting in outward

movement at the round window and vice-versa. As a result of this mechanical stimulation, the cochlear partition reacts in a remarkable way: the different parts of the partition are entrained by different frequency components of the incoming pressure wave. The most basal stretches of the partition are entrained by the highest frequency components and, as one goes toward the apex, the partition becomes entrained by lower and lower frequency components. An intuitive explanation for this phenomenon can be given in terms of path of least resistance. Movement in a system occurs preferentially along the path of least resistance and in the cochlea the path of least resistance happens to be different for the different frequency components of an incoming sound. Indeed, because the cochlear partition gets more and more flexible as one goes toward the apex of the cochlea, more apical places offer less resistance to motion. But, moving the cochlear partition at more apical locations requires moving larger masses of fluid (from the oval window to the point where the movement occurs and back to the round window). The optimal compromise between these two antagonist influences determines the path of least resistance. It is frequency-dependent because of the inertia of the fluid: moving fluid at a higher frequency, i.e. faster, requires more energy.

Thus the cochlear partition appears to operate as a sort of *spatial Fourier analyzer* of the incoming sound, in the sense that the motion at different points of the cochlear partition is related to the amount of energy in different frequency bands in the incoming sound. An important component of Fourier analysis, however, is linearity and the cochlea is known to be nonlinear as it exhibits several phenomenons that cannot arise in linear systems ([243], Chapter 4) . Most of the elements in the physical description of the cochlea we outlined above are well-modeled by Linear Time-Invariant (LTI) differential equations ([243], Chapter 3) and thus cannot be the source of the non-linear phenomenon observed in the cochlea. Only the action of the OHC is not well captured by linear models and they are thought to be the main origin of the non-linear phenomenons observed in the cochlea. The exact mechanism of operation of the OHC is still a matter of debate, but it is generally thought that their main functional role is to input mechanical energy into the cochlear partition in a way that makes it an active sensor with a better sensitivity and a better frequency resolution than it could have as a purely passive sensor, allowing to detect weaker sounds and resolve frequency components better. With respect to sensitivity, in particular, the OHC action allows the ear to detect extremely weak sounds, resulting in a

huge 120 dB dynamic range of intensities to which the ear is responsive. This happens despite the limited 30-35 dB dynamic range of possible cochlear partition motion available to encode these intensities. It is possible thanks to a non-linear compression effect, whereby doubling the intensity of a sound results in most cases in less than a doubling of the corresponding displacement of the cochlear partition [168]. This compression of the dynamic range is the most prominent nonlinearity observed in the cochlea, leading to simple phenomenological models where the cochlea is decomposed into a purely linear part (modeled as an LTI system) followed by a static non-linearity that models the dynamic range compression.

Bibliography

[1] Janet F Werker and Richard C Tees. "Influences on infant speech processing: Toward a new synthesis". In: *Annual review of psychology* 50.1 (1999), pp. 509–535 (cit. on pp. 5, 59, 60, 91).

[2] Patricia K Kuhl, Erica Stevens, Akiko Hayashi, Toshisada Deguchi, Shigeru Kiritani, and Paul Iverson. "Infants show a facilitation effect for native language phonetic perception between 6 and 12 months". In: *Developmental science* 9.2 (2006), F13–F21 (cit. on pp. 5, 91).

[3] Bart De Boer and Patricia K Kuhl. "Investigating the role of infant-directed speech with a computer model". In: *Acoustics Research Letters Online* 4.4 (2003), pp. 129–134 (cit. on pp. 5, 152).

[4] Michael H Coen. "Self-supervised acquisition of vowels in American English". In: *Proc. National Conference On Artificial Intelligence*. 2006 (cit. on pp. 5, 152).

[5] Bruno Gauthier, Rushen Shi, and Yi Xu. "Learning phonetic categories by tracking movements". In: *Cognition* 103.1 (2007), pp. 80–106 (cit. on pp. 5, 152).

[6] Gautam K Vallabha, James L McClelland, Ferran Pons, Janet F Werker, and Shigeaki Amano. "Unsupervised learning of vowel categories from infant-directed speech". In: *Proceedings of the National Academy of Sciences* 104.33 (2007), pp. 13273–13278 (cit. on pp. 5, 152).

[7] Kouki Miyazawa, Hideaki Kikuchi, and Reiko Mazuka. "Unsupervised learning of vowels from continuous speech based on self-organized phoneme acquisition model". In: *Proc. INTERSPEECH*. 2010 (cit. on pp. 5, 152).

[8] Kouki Miyazawa, Hideaki Miura, Hideaki Kikuchi, and Reiko Mazuka. "The Multi Timescale Phoneme Acquisition Model of the Self-Organizing Based on the Dynamic Features". In: *Proc. INTERSPEECH*. 2011 (cit. on pp. 5, 152).

[9] Frans Adriaans and Daniel Swingley. "Distributional learning of vowel categories is supported by prosody in infant-directed speech". In: *Proc. CogSci*. 2012 (cit. on pp. 5, 152).

[10] Caroline Jones, Felicity Meakins, and Shujau Muawiyath. "Learning vowel categories from maternal speech in Gurindji Kriol". In: *Language Learning* 62.4 (2012), pp. 1052–1078 (cit. on pp. 5, 152).

[11] Brian Dillon, Ewan Dunbar, and William Idsardi. "A Single-Stage Approach to Learning Phonological Categories: Insights From Inuktitut". In: *Cognitive Science* 37.2 (2013), pp. 344–377 (cit. on pp. 5, 152).

[12] Naomi H Feldman, Thomas L Griffiths, Sharon Goldwater, and James L Morgan. "A role for the developing lexicon in phonetic category acquisition." In: *Psychological review* 120.4 (2013), p. 751 (cit. on pp. 5, 152).

[13] Heikki Rasilo, Okko Räsänen, and Unto K Laine. "Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion". In: *Speech Communication* 55.9 (2013), pp. 909–931 (cit. on pp. 5, 152).

[14] Stella Frank, Naomi Feldman, and Sharon Goldwater. "Weak semantic context helps phonetic learning in a model of infant language acquisition". In: *Proc. ACL*. 2014 (cit. on pp. 5, 152).

[15] David M Green and John A Swets. "Signal detection theory and psychophysics". In: (1966) (cit. on p. 19).

[16] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. "Theory of classification: A survey of some recent advances". In: *ESAIM: probability and statistics* 9 (2005), pp. 323–375 (cit. on p. 21).

[17] Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. "Consistency of spectral clustering". In: *The Annals of Statistics* (2008), pp. 555–586 (cit. on pp. 23, 33, 36).

[18] Paul Mermelstein. "Distance measures for speech recognition, psychological and instrumental". In: *Pattern recognition and artificial intelligence* 116 (1976), pp. 91–103 (cit. on pp. 26, 78, 91, 97, 98).

[19] Taras K Vintsyuk. "Speech discrimination by dynamic programming". In: *Cybernetics and Systems Analysis* 4.1 (1968), pp. 52–57 (cit. on pp. 26, 133).

[20] Douglas B Paul and Janet M Baker. "The design for the Wall Street Journal-based CSR corpus". In: *Proc. Workshop on Speech and Natural Language*. 1992, pp. 357–362 (cit. on pp. 26, 83, 130).

[21] Trevor J.. Hastie, Robert John Tibshirani, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009 (cit. on p. 29).

[22] Jianbo Shi and Jitendra Malik. "Normalized cuts and image segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8 (2000), pp. 888–905 (cit. on pp. 29, 32).

[23] Hamid K Seifoddini. "Single linkage versus average linkage clustering in machine cells formation applications". In: *Computers & Industrial Engineering* 16.3 (1989), pp. 419–426 (cit. on p. 30).

[24] Lawrence Hubert and Phipps Arabie. "Comparing partitions". In: *Journal of classification* 2.1 (1985), pp. 193–218 (cit. on p. 30).

[25] Tong Zhang. "Statistical behavior and consistency of classification methods based on convex risk minimization". In: *Annals of Statistics* (2004), pp. 56–85 (cit. on p. 32).

[26] David Arthur and Sergei Vassilvitskii. "k-means++: The advantages of careful seeding". In: *Proc. ACM-SIAM symposium on discrete algorithms*. 2007 (cit. on p. 32).

[27] Frédéric Cérou and Arnaud Guyader. "Nearest neighbor classification in infinite dimension". In: *ESAIM: Probability and Statistics* 10 (2006), pp. 340–355 (cit. on p. 33).

[28] Solomon Kullback and Richard A Leibler. "On information and sufficiency". In: *The annals of mathematical statistics* (1951), pp. 79–86 (cit. on pp. 33, 133).

[29] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012 (cit. on p. 34).

[30] David D Lewis. “Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval”. In: *Proc. European Conference on Machine Learning*. 1998 (cit. on p. 36).

[31] Luc Devroye, László Györfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Vol. 31. Springer Science & Business Media, 1996 (cit. on p. 36).

[32] Ohad Shamir and Naftali Tishby. “Stability and model selection in k-means clustering”. In: *Machine learning* 80.2-3 (2010), pp. 213–243 (cit. on p. 36).

[33] Ulrike von Luxburg. “Clustering Stability: An Overview”. In: *Machine Learning* 2.3 (2009), pp. 235–274 (cit. on p. 36).

[34] Michael A Carlin, Samuel Thomas, Aren Jansen, and Hynek Hermansky. “Rapid Evaluation of Speech Representations for Spoken Term Discovery.” In: *Proc. INTERSPEECH*. 2011 (cit. on pp. 55, 57, 65, 68, 71, 111).

[35] Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. “Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline”. In: *Proc. INTERSPEECH*. 2013 (cit. on pp. 55, 58, 68, 71, 100, 151).

[36] Jesse Davis and Mark Goadrich. “The relationship between Precision-Recall and ROC curves”. In: *Proc. ICML*. 2006 (cit. on p. 57).

[37] Thomas Schatz, Xuan-Nga Cao, Anna Kolesnikova, Tomas Bergvelt, Jonathan Wright, and Emmanuel Dupoux. *Articulation Index LSCP LDC2015S12*. https://catalog.ldc.upenn.edu/LDC2015S12. 2015 (cit. on pp. 58, 109).

[38] Maarten Versteegh, Roland Thiolliere, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. “The zero resource speech challenge 2015”. In: *Proc. INTERSPEECH*. 2015 (cit. on pp. 58, 69, 151).

[39] Eleni Kotsoni, Michelle de Haan, and Mark H Johnson. “Categorical perception of facial expressions by 7-month-old infants”. In: *Perception* 30.9 (2001), pp. 1115–1125 (cit. on p. 59).

[40] Scott P Johnson and Erin E Hannon. *Handbook of child psychology and developmental science: Vol. 2, Chap. 3, Perceptual development*. Seventh. John Wiley & Sons Inc, 2015 (cit. on p. 59).

[41] Jean M Mandler and Laraine McDonough. “Concept formation in infancy”. In: *Cognitive development* 8.3 (1993), pp. 291–318 (cit. on p. 59).

[42] Jean M Mandler and Laraine McDonough. “On developing a knowledge base in infancy.” In: *Developmental psychology* 34.6 (1998), p. 1274 (cit. on p. 59).

[43] Paul C Quinn and Mark H Johnson. “Global-Before-Basic Object Categorization in Connectionist Networks and 2-Month-Old Infants”. In: *Infancy* 1.1 (2000), pp. 31–46 (cit. on p. 59).

[44] Birgit Elsner, Susanna Jeschonek, and Sabina Pauen. “Event-related potentials for 7-month-olds? processing of animals and furniture items”. In: *Developmental cognitive neuroscience* 3 (2013), pp. 53–60 (cit. on p. 59).

[45] Michèle Molina, Gretchen A Van de Walle, Kirsten Condry, and Elizabeth S Spelke. “The animate-inanimate distinction in infancy: Developing sensitivity to constraints on human actions”. In: *Journal of cognition and development* 5.4 (2004), pp. 399–426 (cit. on p. 59).

[46] Diane Poulin-Dubois, Anouk Lepage, and Doreen Ferland. “Infants’ concept of animacy”. In: *Cognitive Development* 11.1 (1996), pp. 19–36 (cit. on p. 59).

[47] Andréa Aguiar and Renée Baillargeon. “2.5-month-old infants’ reasoning about when objects should and should not be occluded”. In: *Cognitive psychology* 39.2 (1999), pp. 116–157 (cit. on p. 59).

[48] Andréa Aguiar and Renée Baillargeon. “Developments in young infants’ reasoning about occluded objects”. In: *Cognitive psychology* 45.2 (2002), pp. 267–336 (cit. on p. 59).

[49] Yuyan Luo and Renée Baillargeon. “When the ordinary seems unexpected: evidence for incremental physical knowledge in young infants”. In: *Cognition* 95.3 (2005), pp. 297–328 (cit. on p. 59).

[50] Susan J Hespos, Alissa L Ferry, and Lance J Rips. “Five-month-old infants have different expectations for solids and liquids”. In: *Psychological Science* 20.5 (2009), pp. 603–611 (cit. on p. 59).

[51] Véronique Izard, Coralie Sann, Elizabeth S Spelke, and Arlette Streri. “Newborn infants perceive abstract numbers”. In: *Proceedings of the National Academy of Sciences* 106.25 (2009), pp. 10382–10385 (cit. on p. 59).

[52] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444 (cit. on p. 60).

[53] MI Jordan and TM Mitchell. "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245 (2015), pp. 255–260 (cit. on p. 60).

[54] Aren Jansen, Emmanuel Dupoux, Sharon Goldwater, Mark Johnson, Sanjeev Khudanpur, Kenneth Church, Naomi Feldman, Hynek Hermansky, Florian Metze, Richard Rose, Michael Seltzer, Pascal Clark, Ian McGraw, Balakrishnan Varadarajan, Erin Bennett, Benjamin Borschinger, Justin Chiu, Ewan Dunbar, Abdellah Fourtassi, David Harwath, Chia-ying Lee, Keith Levin, Atta Norouzian, Vijayaditya Peddinti, Rachael Richardson, Thomas Schatz, and Samuel Thomas. "A summary of the 2012 JH CLSP Workshop on zero resource speech technologies and models of early language acquisition". In: *Proc. ICASSP*. 2013 (cit. on pp. 62, 67, 68).

[55] PK Kuhl and AN Meltzoff. "The bimodal perception of speech in infancy". In: *Science* 218.4577 (1982), pp. 1138–1141 (cit. on p. 62).

[56] Michelle L Patterson and Janet F Werker. "Two-month-old infants match phonetic information in lips and voice". In: *Developmental Science* 6.2 (2003), pp. 191–196 (cit. on p. 62).

[57] Jessica Maye, Janet F Werker, and LouAnn Gerken. "Infant sensitivity to distributional information can affect phonetic discrimination". In: *Cognition* 82.3 (2002), B101–B111 (cit. on p. 62).

[58] Jessica Maye, Daniel J Weiss, and Richard N Aslin. "Statistical phonetic learning in infants: Facilitation and feature generalization". In: *Developmental science* 11.1 (2008), pp. 122–134 (cit. on p. 62).

[59] Katherine A Yoshida, Ferran Pons, Jessica Maye, and Janet F Werker. "Distributional phonetic learning at 10 months of age". In: *Infancy* 15.4 (2010), pp. 420–433 (cit. on p. 62).

[60] Tuomas Teinonen, Richard N Aslin, Paavo Alku, and Gergely Csibra. "Visual speech contributes to phonetic learning in 6-month-old infants". In: *Cognition* 108.3 (2008), pp. 850–855 (cit. on p. 62).

[61] Alejandrina Cristia. "Fine-grained variation in caregivers' /s/ predicts their infants' /s/ category". In: *The Journal of the Acoustical Society of America* 129.5 (2011), pp. 3271–3280 (cit. on p. 62).

[62] Emily JH Jones and Jane S Herbert. "The effect of learning experiences and context on infant imitation and generalization". In: *Infancy* 13.6 (2008), pp. 596–619 (cit. on p. 62).

[63] Derek M Houston and Peter W Jusczyk. "Infants' long-term memory for the sound patterns of words and voices." In: *Journal of Experimental Psychology: Human Perception and Performance* 29.6 (2003), p. 1143 (cit. on p. 62).

[64] H Henny Yeung and Janet F Werker. "Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information". In: *Cognition* 113.2 (2009), pp. 234–243 (cit. on p. 62).

[65] Nivedita Mani and Signe Schneider. "Speaker identity supports phonetic category learning." In: *Journal of Experimental Psychology: Human Perception and Performance* 39.3 (2013), p. 623 (cit. on p. 62).

[66] Céline Ngon, Andrew Martin, Emmanuel Dupoux, Dominique Cabrol, Michel Dutat, and Sharon Peperkamp. "(Non) words,(non) words,(non) words: evidence for a protolexicon during the first year of life". In: *Developmental science* 16.1 (2013), pp. 24–34 (cit. on p. 62).

[67] Kathy Hirsh-Pasek, Deborah G Kemler Nelson, Peter W Jusczyk, Kimberly Wright Cassidy, Benjamin Druss, and Lori Kennedy. "Clauses are perceptual units for young infants". In: *Cognition* 26.3 (1987), pp. 269–286 (cit. on p. 62).

[68] Peter W Jusczyk, Kathy Hirsh-Pasek, Deborah G Kemler Nelson, Lori J Kennedy, Amanda Woodward, and Julie Piwoz. "Perception of acoustic correlates of major phrasal units by young infants". In: *Cognitive psychology* 24.2 (1992), pp. 252–293 (cit. on p. 62).

[69] James Myers, Peter W Jusczyk, Deborah G Kemler Nelson, Jan Charles-Luce, Amanda L Woodward, and Kathryn Hirsh-Pasek. "Infants' sensitivity to word boundaries in fluent speech". In: *Journal of Child Language* 23.01 (1996), pp. 1–30 (cit. on p. 62).

[70] Naomi H Feldman, Emily B Myers, Katherine S White, Thomas L Griffiths, and James L Morgan. "Word-level information influences phonetic learning in adults and infants". In: *Cognition* 127.3 (2013), pp. 427–438 (cit. on p. 62).

[71] Elika Bergelson and Daniel Swingley. "At 6–9 months, human infants know the meanings of many common nouns". In: *Proceedings of the National Academy of Sciences* 109.9 (2012), pp. 3253–3258 (cit. on p. 62).

[72] Elika Bergelson and Daniel Swingley. "Early Word Comprehension in Infants: Replication and Extension". In: *Language Learning and Development* ahead-of-print (2014), pp. 1–12 (cit. on p. 62).

[73] H Henny Yeung, Lawrence M Chen, and Janet F Werker. "Referential labeling can facilitate phonetic learning in infancy". In: *Child development* 85.3 (2014), pp. 1036–1049 (cit. on p. 62).

[74] H Henny Yeung and Thierry Nazzi. "Object labeling influences infant phonetic learning and generalization". In: *Cognition* 132.2 (2014), pp. 151–163 (cit. on p. 62).

[75] Anne Fernald. "Four-month-old infants prefer to listen to motherese". In: *Infant behavior and development* 8.2 (1985), pp. 181–195 (cit. on p. 62).

[76] Michael Scaife and Jerome S Bruner. "The capacity for joint visual attention in the infant." In: *Nature* (1975) (cit. on p. 62).

[77] George Butterworth. "Pointing is the royal road to language for babies". In: *Pointing: Where language, culture, and cognition meet* (2003), pp. 9–33 (cit. on p. 62).

[78] George Butterworth. "Joint visual attention in infancy". In: *Theories of infant development* (2004), pp. 317–354 (cit. on p. 62).

[79] Patricia K Kuhl, Feng-Ming Tsao, and Huei-Mei Liu. "Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning". In: *Proceedings of the National Academy of Sciences* 100.15 (2003), pp. 9096–9101 (cit. on p. 62).

[80] Douglas H Whalen, Andrea G Levitt, and Qi Wang. "Intonational differences between the reduplicative babbling of French-and English-learning infants". In: *Journal of Child Language* 18.03 (1991), pp. 501–516 (cit. on p. 62).

[81] Bénédicte de Boysson-Bardies, Laurent Sagart, and Catherine Durand. "Discernible differences in the babbling of infants according to target language". In: *Journal of child language* 11.01 (1984), pp. 1–15 (cit. on p. 62).

[82] H Henny Yeung and Janet F Werker. "Lip movements affect infants' audiovisual speech perception". In: *Psychological Science* 24.5 (2013), pp. 603–612 (cit. on p. 62).

[83] Patricia K Kuhl and Andrew N Meltzoff. "Infant vocalizations in response to speech: Vocal imitation and developmental change". In: *The journal of the Acoustical Society of America* 100.4 (1996), pp. 2425–2438 (cit. on p. 62).

[84] Michael H Goldstein and Jennifer A Schwade. "Social feedback to infants' babbling facilitates rapid phonological learning". In: *Psychological Science* 19.5 (2008), pp. 515–523 (cit. on p. 62).

[85] Thierry Nazzi and Franck Ramus. "Perception and acquisition of linguistic rhythm by infants". In: *Speech Communication* 41.1 (2003), pp. 233–243 (cit. on p. 62).

[86] Aren Jansen and Kenneth Church. "Towards Unsupervised Training of Speaker Independent Acoustic Models." In: *Proc. INTERSPEECH*. 2011 (cit. on pp. 66, 67, 69, 71).

[87] Aren Jansen, Samuel Thomas, and Hynek Hermansky. "Weak top-down constraints for unsupervised acoustic model training." In: *Proc. ICASSP*. 2013 (cit. on pp. 67, 69–71).

[88] Aren Jansen, Samuel Thomas, and Hynek Hermansky. "Intrinsic Spectral Analysis for Zero and High Resource Speech Recognition." In: *Proc. INTERSPEECH*. 2012 (cit. on pp. 67, 71).

[89] Hynek Hermansky and David J Broad. "The effective second formant F2'and the vocal tract front-cavity". In: *Proc. ICASSP*. 1989 (cit. on p. 67).

[90] Chia-ying Lee and James Glass. "A nonparametric Bayesian approach to acoustic model discovery". In: *Proc. ACL*. 2012 (cit. on pp. 67, 68, 71).

[91] Herman Kamper, Weiran Wang, and Karen Livescu. "Deep convolutional acoustic word embeddings using word-pair side information". In: *arXiv preprint arXiv:1510.01032* (2015) (cit. on pp. 68, 70).

[92] Li Deng, Geoffrey Hinton, and Brian Kingsbury. "New types of deep neural network learning for speech recognition and related applications: An overview". In: *Proc. ICASSP*. 2013 (cit. on p. 68).

[93] László Tóth. "Combining time-and frequency-domain convolution in convolutional neural network-based phone recognition". In: *Proc. ICASSP*. 2014 (cit. on p. 68).

[94] David Snyder, Daniel Garcia-Romero, and Daniel Povey. "Time delay deep neural network-based universal background models for speaker recognition". In: *Proc. ASRU*. 2015 (cit. on p. 68).

[95] Karthika Vijayan, Pappagari Raghavendra Reddy, and K Sri Rama Murty. "Significance of analytic phase of speech signals in speaker verification". In: *Speech Communication* 81 (2016), pp. 54–71 (cit. on pp. 68, 151).

[96] Thomas Schatz, Vijayaditya Peddinti, Xuan-Nga Cao, Francis R Bach, Hynek Hermansky, and Emmanuel Dupoux. "Evaluating speech features with the minimal-pair ABX task (II): resistance to noise." In: *Proc. INTERSPEECH*. 2014 (cit. on pp. 68, 71, 100, 151).

[97] Gabriel Synnaeve, Thomas Schatz, and Emmanuel Dupoux. "Phonetics embedding learning with side information". In: *Proc. Spoken Language Technology Workshop*. 2014 (cit. on pp. 69–71, 151).

[98] Maarten Versteegh, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. "The Zero Resource Speech Challenge 2015: Proposed Approaches and Results". In: *Procedia Computer Science* 81 (2016), pp. 67–72 (cit. on pp. 69, 151).

[99] Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. "The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability". In: *Speech Communication* 45.1 (2005), pp. 89–95 (cit. on pp. 69, 130).

[100] Nic J De Vries, Marelie H Davel, Jaco Badenhorst, Willem D Basson, Febe De Wet, Etienne Barnard, and Alta De Waal. "A smartphone-based ASR data collection tool for under-resourced languages". In: *Speech communication* 56 (2014), pp. 119–131 (cit. on p. 69).

[101] Roland Thiollière, Ewan Dunbar, Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux. "A Hybrid Dynamic Time Warping-Deep Neural Network Architecture for Unsupervised Acoustic Modeling". In: *Proc. INTERSPEECH*. 2015 (cit. on pp. 70, 71, 151).

[102] Daniel Renshaw, Herman Kamper, Aren Jansen, and Sharon Goldwater. "A comparison of neural network methods for unsupervised representation learning on the Zero Resource Speech Challenge". In: *Proc. INTERSPEECH*. 2015 (cit. on pp. 70, 71, 151).

[103] Neil Zeghidour, Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux. "A deep scattering spectrum - Deep Siamese network pipeline for unsupervised acoustic modeling". In: *Proc. ICASSP*. 2016 (cit. on pp. 70, 71, 151).

[104] Joakim Andén and Stéphane Mallat. "Deep scattering spectrum". In: *IEEE Transactions on Signal Processing* 62.16 (2014), pp. 4114–4128 (cit. on pp. 70, 103, 106, 107, 121, 166).

[105] Leonardo Badino, Alessio Mereta, and Lorenzo Rosasco. "Discovering Discrete Subword Units with Binarized Autoencoders and Hidden-Markov-Model Encoders". In: *Proc. INTERSPEECH*. 2015 (cit. on pp. 70, 71, 151).

[106] Hongjie Chen, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li. "Parallel Inference of Dirichlet Process Gaussian Mixture Models for Unsupervised Acoustic Modeling: A Feasibility Study". In: *Proc. INTERSPEECH*. 2015 (cit. on pp. 70, 71, 151).

[107] Michael Heck, Sakriani Sakti, and Satoshi Nakamura. "Unsupervised Linear Discriminant Analysis for Supporting DPGMM Clustering in the Zero Resource Scenario". In: *Procedia Computer Science* 81 (2016), pp. 73–79 (cit. on pp. 71, 151).

[108] Cheng-Tao Chung, Cheng-Yu Tsai, Hsiang-Hung Lu, Chia-Hsiang Liu, Hung-yi Lee, and Lin-shan Lee. "An Iterative Deep Learning Framework for Unsupervised Discovery of Speech Features and Linguistic Units With Applications on Spoken Term Detection". In: *Proc. ASRU*. 2015 (cit. on pp. 71, 151).

[109] NA Macmillan and CD Creelman. "Detection Theory: A User's Guide Lawrence Erlbaum Associates". In: *New York* (2005) (cit. on pp. 73, 76).

[110] Yao Yao, Sam Tilsen, Ronald L Sprouse, and Keith Johnson. "Automated measurement of vowel formants in the buckeye corpus". In: *UC Berkeley Phonology Lab Annual Report* (2010) (cit. on p. 78).

[111] Neville Ryant, Jiahong Yuan, and Mark Liberman. "Automating phonetic measurement: The case of voice onset time". In: *Proc. Meetings on Acoustics*. 2013 (cit. on p. 78).

[112] Hynek Hermansky. "Perceptual linear predictive (PLP) analysis of speech". In: *The Journal of the Acoustical Society of America* 87 (1990), pp. 1738–1752 (cit. on pp. 78, 91, 97, 98).

[113] Ralf Schluter, L Bezrukov, Hannes Wagner, and Hermann Ney. "Gammatone features and feature combination for large vocabulary speech recognition". In: *Proc. ICASSP*. 2007 (cit. on p. 78).

[114] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 82–97 (cit. on p. 78).

[115] Andrew Martin, Thomas Schatz, Maarten Versteegh, Kouki Miyazawa, Reiko Mazuka, Emmanuel Dupoux, and Alejandrina Cristia. "Mothers Speak Less Clearly to Infants Than to Adults A Comprehensive Test of the Hyperarticulation Hypothesis". In: *Psychological science* 26.3 (2015), pp. 341–347 (cit. on pp. 78, 79, 120, 151).

[116] Alejandrina Cristia. "Input to Language: The Phonetics and Perception of Infant-Directed Speech". In: *Language and Linguistics Compass* 7.3 (2013), pp. 157–170 (cit. on p. 79).

[117] Yosuke Igarashi, Ken'ya Nishikawa, Kuniyoshi Tanaka, and Reiko Mazuka. "Phonological theory informs the analysis of intonational exaggeration in Japanese infant-directed speech". In: *The Journal of the Acoustical Society of America* 134.2 (2013), pp. 1283–1294 (cit. on p. 79).

[118] R Mazuka, Y Igarashi, and K Nishikawa. *Input for learning Japanese: RIKEN Japanese Mother-Infant Conversation Corpus. The Institute of Electronics, Information and Communication Engineers*. Tech. rep. Technical Report TL2006-16 (2006-07): 11–15, 2006 (cit. on p. 79).

[119] Bob McMurray, Kristine A Kovack-Lesh, Dresden Goodwin, and William McEchron. "Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence?" In: *Cognition* 129.2 (2013), pp. 362–378 (cit. on p. 81).

[120] Titia Benders. "Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent". In: *Infant Behavior and Development* 36.4 (2013), pp. 847–862 (cit. on p. 81).

[121] Nairán Ramírez-Esparza, Adrián García-Sierra, and Patricia K Kuhl. "Look who's talking: speech style and social context in language input to infants are linked to concurrent and future speech development". In: *Developmental science* 17.6 (2014), pp. 880–891 (cit. on p. 81).

[122] Adriana Weisleder and Anne Fernald. "Talking to children matters early language experience strengthens processing and builds vocabulary". In: *Psychological Science* 24.11 (2013), pp. 2143–2152 (cit. on p. 81).

[123] Peter Ladefoged and Keith Johnson. *A course in phonetics*. Nelson Education, 2014 (cit. on p. 81).

[124] Kikuo Maekawa. "Corpus of Spontaneous Japanese: Its design and evaluation". In: *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*. 2003 (cit. on pp. 83, 130).

[125] R Smits. "Evidence for hierarchical categorization of coarticulated phonemes." In: *Journal of experimental psychology. Human perception and performance* 27.5 (2001), p. 1145 (cit. on p. 86).

[126] Florian Hönig, Georg Stemmer, Christian Hacker, and Fabio Brugnara. "Revising Perceptual Linear Prediction (PLP)." In: *Proc. INTERSPEECH*. 2005 (cit. on pp. 91, 97, 98).

[127] Peter D Eimas, Einar R Siqueland, Peter W Jusczyk, and James Vigorito. "Speech perception in infants". In: *Science* 171 (1971), pp. 303–306 (cit. on pp. 91, 151).

[128] Peter D Eimas. "Auditory and linguistic processing of cues for place of articulation by infants". In: *Perception & Psychophysics* 16.3 (1974), pp. 513–521 (cit. on pp. 91, 151).

[129] Peter D Eimas. "Auditory and phonetic coding of the cues for speech: Discrimination of the [rl] distinction by young infants". In: *Perception & Psychophysics* 18.5 (1975), pp. 341–347 (cit. on pp. 91, 151).

[130] Lynn A Streeter. "Language perception of 2-mo-old infants shows effects of both innate mechanisms and experience." In: *Nature* (1976) (cit. on p. 91).

[131] Sandra E Trehub. "The discrimination of foreign speech contrasts by infants and adults". In: *Child development* (1976), pp. 466–472 (cit. on p. 91).

[132] Richard N Aslin, David B Pisoni, Beth L Hennessy, and Alan J Perey. "Discrimination of voice onset time by human infants: New findings and implications for the effects of early experience". In: *Child development* 52.4 (1981), p. 1135 (cit. on p. 91).

[133] Patricia K Kuhl and James D Miller. "Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants". In: *Perception & Psychophysics* 31.3 (1982), pp. 279–292 (cit. on p. 91).

[134] Janet F Werker and Richard C Tees. "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life". In: *Infant behavior and development* 7.1 (1984), pp. 49–63 (cit. on p. 91).

[135] Josiane Bertoncini, Ranka Bijeljac-Babic, Sheila E Blumstein, and Jacques Mehler. "Discrimination in neonates of very short CVs". In: *The Journal of the Acoustical Society of America* 82.1 (1987), pp. 31–37 (cit. on p. 91).

[136] Josiane Bertoncini, Ranka Bijeljac-Babic, Peter W Jusczyk, Lori J Kennedy, and Jacques Mehler. "An investigation of young infants' perceptual representations of speech sounds." In: *Journal of experimental psychology: General* 117.1 (1988), p. 21 (cit. on p. 91).

[137] Linda Polka and Janet F Werker. "Developmental changes in perception of nonnative vowel contrasts." In: *Journal of Experimental Psychology: Human Perception and Performance* 20.2 (1994), p. 421 (cit. on p. 91).

[138] Teruaki Tsushima, Osamu Takizawa, Midori Sasaki, Satoshi Shiraki, Kanae Nishi, Morio Kohno, Paula Menyuk, and Catherine T Best. "Discrimination of English/rl/and/wy/by Japanese infants at 6-12 months: language-specific developmental changes in speech perception abilities." In: *Proc. ICSLP*. 1994 (cit. on p. 91).

[139] Anu Kujala, Minna Huotilainen, Merja Hotakainen, Mietta Lennes, Lauri Parkkonen, Vineta Fellman, and Risto Näätänen. "Speech-sound discrimination in neonates as measured with MEG". In: *Neuroreport* 15.13 (2004), pp. 2089–2092 (cit. on p. 91).

[140] Patricia K Kuhl. "Theoretical contributions of tests on animals to the special-mechanisms debate in speech." In: *Experimental biology* 45.3 (1985), pp. 233–265 (cit. on p. 92).

[141] Randy L Diehl, Andrew J Lotto, and Lori L Holt. "Speech perception". In: *Annu. Rev. Psychol.* 55 (2004), pp. 149–179 (cit. on p. 92).

[142] Ruth Tincoff, Marc Hauser, Fritz Tsao, Geertrui Spaepen, Franck Ramus, and Jacques Mehler. "The role of speech rhythm in language discrimination: further tests with a non-human primate". In: *Developmental science* 8.1 (2005), pp. 26–35 (cit. on p. 92).

[143] Lori L Holt and Andrew J Lotto. "Speech perception within an auditory cognitive science framework". In: *Current directions in psychological science* 17.1 (2008), pp. 42–46 (cit. on p. 92).

[144] Patricia K Kuhl and James D Miller. "Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants". In: *Science* 190.4209 (1975), pp. 69–72 (cit. on p. 92).

[145] Patricia K Kuhl and James D Miller. "Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli". In: *The Journal of the Acoustical Society of America* 63.3 (1978), pp. 905–917 (cit. on p. 92).

[146] Patricia K Kuhl. "Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories". In: *The Journal of the Acoustical Society of America* 70.2 (1981), pp. 340–349 (cit. on p. 92).

[147] Patricia K Kuhl and Denise M Padden. "Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques". In: *Perception & Psychophysics* 32.6 (1982), pp. 542–550 (cit. on p. 92).

[148] Patricia K Kuhl and Denise M Padden. "Enhanced discriminability at the phonetic boundaries for the place feature in macaques". In: *The Journal of the Acoustical Society of America* 73.3 (1983), pp. 1003–1010 (cit. on p. 92).

[149] W Tecumseh Fitch. "The evolution of speech: a comparative review". In: *Trends in cognitive sciences* 4.7 (2000), pp. 258–267 (cit. on p. 93).

[150] Philip Lieberman. "The evolution of human speech". In: *Current Anthropology* 48.1 (2007), pp. 39–66 (cit. on p. 93).

[151] Ann B Butler and William Hodos. *Comparative vertebrate neuroanatomy: evolution and adaptation*. John Wiley & Sons, 2005 (cit. on p. 93).

[152] Jan Schnupp, Israel Nelken, and Andrew King. *Auditory neuroscience: Making sense of sound*. MIT Press, 2011 (cit. on p. 93).

[153] Evan C Smith and Michael S Lewicki. "Efficient auditory coding". In: *Nature* 439.7079 (2006), pp. 978–982 (cit. on p. 93).

[154] Jenny R Saffran, Janet F Werker, and Lynne A Werner. "The infant's auditory world: Hearing, speech, and the beginnings of language". In: *Handbook of child psychology* (2006) (cit. on p. 94).

[155] Daniel P. W. Ellis. *PLP and RASTA (and MFCC, and inversion) in Matlab*. http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat. 2005 (cit. on p. 96).

[156] Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky. "Spectro-temporal features for automatic speech recognition using linear prediction in spectral domain". In: *Proc. European Signal Processing Conference*. 2008 (cit. on pp. 96, 97, 100).

[157] Marios Athineos and Daniel PW Ellis. "Frequency-domain linear prediction for temporal features". In: *Proc. ASRU*. 2003 (cit. on pp. 97, 100).

[158] Hynek. Hermansky and Nelson Morgan. "RASTA processing of speech". In: *IEEE Transactions on Speech and Audio Processing* 2.4 (1994), pp. 578–589 (cit. on pp. 98, 107).

[159] Mead C Killion. "Revised estimate of minimum audible pressure: Where is the "missing 6 dB"". In: *The Journal of the Acoustical Society of America* 63.5 (1978), pp. 1501–1508 (cit. on pp. 100, 101).

[160] Yoiti Suzuki and Hisashi Takeshima. "Equal-loudness-level contours for pure tones". In: *The Journal of the Acoustical Society of America* 116.2 (2004), pp. 918–933 (cit. on pp. 101, 102).

[161] Gunnar Fant, Johan Liljencrants, and Qi-guang Lin. “A four-parameter model of glottal flow”. In: *STL-QPSR* 4.1985 (1985), pp. 1–13 (cit. on p. 103).

[162] Jody Kreiman, Bruce R Gerratt, and Norma Antoñanzas-Barroso. “Measures of the glottal source spectrum”. In: *Journal of speech, language, and hearing research* 50.3 (2007), pp. 595–610 (cit. on pp. 101, 103).

[163] Brian CJ Moore and Brian R Glasberg. “A revised model of loudness perception applied to cochlear hearing loss”. In: *Hearing research* 188.1 (2004), pp. 70–88 (cit. on p. 101).

[164] Denis Byrne, Harvey Dillon, Khanh Tran, Stig Arlinger, Keith Wilbraham, Robyn Cox, Bjorn Hagerman, Raymond Hetu, Joseph Kei, C Lui, et al. “An international comparison of long-term average speech spectra”. In: *The Journal of the Acoustical Society of America* 96.4 (1994), pp. 2108–2120 (cit. on p. 101).

[165] Gunnar Fant. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. Vol. 2. Walter de Gruyter, 1971 (cit. on pp. 101, 103, 108, 109).

[166] M Mohan Sondhi and Juergen Schroeter. “Speech production models and their digital implementations”. In: *The Digital Signal Processing Handbook* (1997) (cit. on pp. 103, 108).

[167] Shihab Shamma and Christian Lorenzi. “On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system”. In: *The Journal of the Acoustical Society of America* 133.5 (2013), pp. 2818–2833 (cit. on p. 103).

[168] Graeme K Yates, Ian M Winter, and Donald Robertson. “Basilar membrane nonlinearity determines auditory nerve rate-intensity functions and cochlear dynamic range”. In: *Hearing research* 45.3 (1990), pp. 203–219 (cit. on pp. 104, 169).

[169] Charles L Philips, John M Parr, and E Riskin. *Signals, systems, and transforms*. Prentice Hall, 1995 (cit. on p. 104).

[170] E De Boer and HR De Jongh. “On cochlear encoding: Potentialities and limitations of the reverse-correlation technique”. In: *The Journal of the Acoustical Society of America* 63.1 (1978), pp. 115–135 (cit. on p. 104).

[171] JO Pickles. *An introduction to the physiology of hearing*. Emerald Group Publishing, 2008 (cit. on p. 105).

[172] BCJ Moore. *An introduction to the psychology of hearing)*. Emerald Group Publishing, 2004 (cit. on pp. 105, 106).

[173] RD Patterson, Ian Nimmo-Smith, John Holdsworth, and Peter Rice. "An efficient auditory filterbank based on the gammatone function". In: *Proc. Meeting of the IOC Speech Group on Auditory Modelling at RSRE*. 1987 (cit. on p. 105).

[174] Eberhard Zwicker and Ernst Terhardt. "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency". In: *The Journal of the Acoustical Society of America* 68.5 (1980), pp. 1523–1525 (cit. on p. 106).

[175] Stanley Smith Stevens, John Volkmann, and Edwin B Newman. "A scale for the measurement of the psychological magnitude pitch". In: *The Journal of the Acoustical Society of America* 8.3 (1937), pp. 185–190 (cit. on p. 106).

[176] Donald D Greenwood. "A cochlear frequency-position function for several species - 29 years later". In: *The Journal of the Acoustical Society of America* 87.6 (1990), pp. 2592–2605 (cit. on p. 106).

[177] Brian CJ Moore and Brian R Glasberg. "A revision of Zwicker's loudness model". In: *Acta Acustica united with Acustica* 82.2 (1996), pp. 335–345 (cit. on p. 106).

[178] Josh H McDermott and Eero P Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis". In: *Neuron* 71.5 (2011), pp. 926–940 (cit. on p. 106).

[179] PX Joris, CE Schreiner, and A Rees. "Neural processing of amplitude-modulated sounds". In: *Physiological reviews* 84.2 (2004), pp. 541–577 (cit. on p. 106).

[180] Robert V Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid. "Speech recognition with primarily temporal cues". In: *Science* 270.5234 (1995), pp. 303–304 (cit. on p. 106).

[181] Zachary M Smith, Bertrand Delgutte, and Andrew J Oxenham. "Chimaeric sounds reveal dichotomies in auditory perception". In: *Nature* 416.6876 (2002), pp. 87–90 (cit. on p. 106).

[182] Steven M Schimmel and Les E Atlas. "Coherent Envelope Detection for Modulation Filtering of Speech." In: *Proc. ICASSP*. 2005 (cit. on pp. 106, 121).

[183] Richard E Turner and Maneesh Sahani. "Demodulation as probabilistic inference". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.8 (2011), pp. 2398–2411 (cit. on pp. 106, 121).

[184] Gregory Sell and Malcolm Slaney. "Solving demodulation as an optimization problem". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.8 (2010), pp. 2051–2066 (cit. on pp. 106, 121).

[185] Taishih Chi, Powen Ru, and Shihab A Shamma. "Multiresolution spectrotemporal analysis of complex sounds". In: *The Journal of the Acoustical Society of America* 118.2 (2005), pp. 887–906 (cit. on pp. 106, 121).

[186] Muhammad SA Zilany, Ian C Bruce, Paul C Nelson, and Laurel H Carney. "A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics". In: *The Journal of the Acoustical Society of America* 126.5 (2009), pp. 2390–2412 (cit. on p. 107).

[187] Brian CJ Moore. "Temporal integration and context effects in hearing". In: *Journal of Phonetics* 31.3 (2003), pp. 563–574 (cit. on p. 107).

[188] Eberhard Zwicker. "?Negative afterimage? in hearing". In: *The Journal of the Acoustical Society of America* 36.12 (1964), pp. 2413–2415 (cit. on p. 107).

[189] Michael Unser. "On the approximation of the discrete Karhunen-Loeve transform for stationary processes". In: *Signal Processing* 7.3 (1984), pp. 231–249 (cit. on p. 108).

[190] Anil K Jain. "A fast Karhunen-Loeve transform for a class of random processes". In: *NASA STI/Recon Technical Report A* 76 (1976), p. 42860 (cit. on p. 108).

[191] Anil K Jain. "A sinusoidal family of unitary transforms". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 4 (1979), pp. 356–365 (cit. on p. 108).

[192] Peter F MacNeilage. "Motor control of serial ordering of speech." In: *Psychological review* 77.3 (1970), p. 182 (cit. on p. 109).

[193] Ramdas Kumaresan and Ashwin Rao. "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications". In: *The Journal of the Acoustical Society of America* 105.3 (1999), pp. 1912–1924 (cit. on p. 121).

[194] Chanwoo Kim and Richard M Stern. "Power-normalized cepstral coefficients (PNCC) for robust speech recognition". In: *Proc. ICASSP*. 2012 (cit. on p. 121).

[195] Anton Jansen and Partha Niyogi. "Intrinsic spectral analysis". In: *IEEE Transactions on Signal Processing* 61.7 (2013), pp. 1698–1710 (cit. on p. 121).

[196] Enrique A Lopez-Poveda. "Spectral processing by the peripheral auditory system: facts and models." In: *International review of neurobiology* 70 (2004), pp. 7–48 (cit. on p. 121).

[197] Morten L Jepsen, Stephan D Ewert, and Torsten Dau. "A computational model of human auditory signal processing and perception". In: *The Journal of the Acoustical Society of America* 124.1 (2008), pp. 422–438 (cit. on p. 121).

[198] Winifred Strange. *Speech perception and linguistic experience: Issues in cross-language research*. York Press, 1995 (cit. on pp. 123, 134).

[199] Anne Cutler. *Native listening: Language experience and the recognition of spoken words*. Mit Press, 2012 (cit. on p. 123).

[200] Hiromu Goto. "Auditory perception by normal Japanese adults of the sounds "L" and "R"". In: *Neuropsychologia* 9.3 (1971), pp. 317–323 (cit. on pp. 124, 129, 134, 138).

[201] Kuniko Miyawaki, James J Jenkins, Winifred Strange, Alvin M Liberman, Robert Verbrugge, and Osamu Fujimura. "An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English". In: *Perception & Psychophysics* 18.5 (1975), pp. 331–340 (cit. on pp. 124, 129, 134, 138).

[202] Catherine T Best. "The emergence of native-language phonological influences in infants: A perceptual assimilation model". In: *The development of speech perception: The transition from speech sounds to spoken words* 167 (1994), p. 224 (cit. on pp. 124, 145, 146).

[203] Catherine T Best. "A Direct Realist View of Cross-Language Speech Perception". In: *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (1995) (cit. on pp. 124, 145, 146).

[204] Catherine T Best, Gerald W McRoberts, and Nomathemba M Sithole. "Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants." In: *Journal of Experimental Psychology: Human perception and performance* 14.3 (1988), p. 345 (cit. on pp. 124, 145, 146).

[205] James E Flege. "Second language speech learning: Theory, findings, and problems". In: *Speech perception and linguistic experience: Issues in cross-language research* (1995), pp. 233–277 (cit. on pp. 125, 145).

[206] James E Flege. "Age of learning and second language speech". In: *Second language acquisition and the critical period hypothesis* (1999), pp. 101–131 (cit. on pp. 125, 145).

[207] James Emil Flege. "Assessing constraints on second-language segmental production and perception". In: *Phonetics and phonology in language comprehension and production: Differences and similarities* 6 (2003), pp. 319–355 (cit. on pp. 125, 145).

[208] Patricia K Kuhl and Paul Iverson. "Chapter 4: Linguistic Experience and the "Perceptual Magnet Effect"". In: *Speech perception and linguistic experience: Issues in cross-language research* (1995), pp. 121–154 (cit. on pp. 125, 145).

[209] Patricia K Kuhl, Barbara T Conboy, Sharon Coffey-Corina, Denise Padden, Maritza Rivera-Gaxiola, and Tobey Nelson. "Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e)". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 363.1493 (2008), pp. 979–1000 (cit. on pp. 125, 126, 145).

[210] Naomi H Feldman, Thomas L Griffiths, and James L Morgan. "The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference." In: *Psychological review* 116.4 (2009), p. 752 (cit. on p. 126).

[211] Laurent Bonnasse-Gahot and Jean-Pierre Nadal. "Neural coding of categories: information efficiency and optimal population codes". In: *Journal of computational neuroscience* 25.1 (2008), pp. 169–187 (cit. on p. 127).

[212] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297 (cit. on p. 127).

[213] Winifred Strange, Miwako Hisagi, Reiko Akahane-Yamada, and Rieko Kubo. "Cross-language perceptual similarity predicts categorial discrimination of American vowels by naive Japanese listeners". In: *The Journal of the Acoustical Society of America* 130.4 (2011), EL226–EL231 (cit. on pp. 127, 135, 143).

[214] Winifred Strange, Ocke-Schwen Bohn, Sonja A Trent, and Kanae Nishi. "Acoustic and perceptual similarity of North German and American English vowels". In: *The Journal of the Acoustical Society of America* 115.4 (2004), pp. 1791–1807 (cit. on pp. 128, 129, 144, 146).

[215] Jian Gong, Martin Cooke, and ML Garcia Lecumberri. "Towards a quantitative model of Mandarin Chinese perception of English consonants". In: *Proc. NewSounds 2010* (2010) (cit. on pp. 128, 129, 146).

[216] Terry L Gottfried. "Effects of consonant context on the perception of French vowels". In: *Journal of Phonetics* 12.2 (1984), pp. 91–114 (cit. on pp. 129, 134).

[217] Arthur S Abramson and Leigh Lisker. "Discriminability along the voicing continuum: Cross-language tests". In: *Proc. International Congress of Phonetic Sciences*. 1970 (cit. on p. 129).

[218] Connie K So and Catherine T Best. "Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences". In: *Language and speech* 53.2 (2010), pp. 273–293 (cit. on p. 129).

[219] Tanja Schultz. "Globalphone: a multilingual speech and text database developed at karlsruhe university." In: *Proc. INTERSPEECH*. 2002 (cit. on p. 130).

[220] Ngoc Thang Vu and Tanja Schultz. "Vietnamese large vocabulary continuous speech recognition". In: *Proc. ASRU*. 2009 (cit. on p. 131).

[221] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, et al. "The Kaldi speech recognition toolkit". In: *Proc. Workshop on Automatic Speech Recognition and Understanding*. 2011 (cit. on p. 131).

[222] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur. "A pitch extraction algorithm tuned for automatic speech recognition". In: *Proc. ICASSP*. 2014 (cit. on p. 131).

[223] Naomi Ogasawara. "Acoustic Analysis of Voice-Onset Time in Taiwan Mandarin and Japanese". In: *Concentric: Studies in Linguistics* 37.2 (2011), pp. 155–178 (cit. on pp. 134, 140).

[224] Naomi Ogasawara. "Production and perception of voice onset time cues in spoken Japanese and Taiwan Mandarin." In: *The Journal of the Acoustical Society of America* 129.4 (2011), pp. 2419–2419 (cit. on p. 134).

[225] Yuh-Shiow Lee, Douglas A Vakoch, and Lee H Wurm. "Tone perception in Cantonese and Mandarin: A cross-linguistic comparison". In: *Journal of Psycholinguistic Research* 25.5 (1996), pp. 527–542 (cit. on pp. 135, 141).

[226] Wen-Shuan Chiao, Baris Kabak, and Bettina Braun. "When more is less: Non-native perception of level tone contrasts". In: *Proc. Psycholinguistic Representation of Tone Conference*. 2011 (cit. on p. 135).

[227] Miwako Hisagi, Valerie L Shafer, Winifred Strange, and Elyse S Sussman. "Perception of a Japanese vowel length contrast by Japanese and American English listeners: Behavioral and electrophysiological measures". In: *Brain research* 1360 (2010), pp. 89–105 (cit. on pp. 135, 142).

[228] Miwako Hisagi and Winifred Strange. "Perception of Japanese temporally-cued contrasts by American English listeners". In: *Language and Speech* 54.2 (2011), pp. 241–264 (cit. on pp. 135, 142).

[229] Gisela Jia, Winifred Strange, Yanhong Wu, Julissa Collado, and Qi Guan. "Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure)". In: *The Journal of the Acoustical Society of America* 119.2 (2006), pp. 1118–1130 (cit. on pp. 135, 143).

[230] Janne Pylkkönen and Mikko Kurimo. "Duration modeling techniques for continuous speech recognition." In: *Proc. INTERSPEECH*. 2004 (cit. on p. 143).

[231] Gabriel Synnaeve and Emmanuel Dupoux. "Weakly supervised multi-embeddings learning of acoustic models". In: *arXiv preprint arXiv:1412.6645* (2014) (cit. on p. 151).

[232] Pappagari Raghavendra Reddy, Karthika Vijayan, and K Sri Rama Murty. "Analysis of features from analytic representation of speech using MP-ABX measures". In: *Proc. INTERSPEECH*. 2015 (cit. on p. 151).

[233] Gabriel Synnaeve and Emmanuel Dupoux. "A Temporal Coherence Loss Function for Learning Unsupervised Acoustic Embeddings". In: *Procedia Computer Science* 81 (2016), pp. 95–100 (cit. on p. 151).

[234] M.J. Carbajal, R. Fér, and E. Dupoux. "Modeling language discrimination in infants using i-vector representations". In: *Proc. CogSci*. 2016 (submitted) (cit. on p. 151).

[235] Vladimir S Korolyuk and YV Borovskich. *Theory of U-statistics*. Vol. 273. Springer Science & Business Media, 2013 (cit. on pp. 153, 154).

[236] CB Bell, David Blackwell, and Leo Breiman. "On the completeness of order statistics". In: *The Annals of Mathematical Statistics* (1960), pp. 794–797 (cit. on p. 154).

[237] Johann Pfanzagl. *Parametric statistical theory*. Walter de Gruyter, 1994 (cit. on p. 155).

[238] Wassily Hoeffding. "Probability inequalities for sums of bounded random variables". In: *Journal of the American statistical association* 58.301 (1963), pp. 13–30 (cit. on p. 158).

[239] Miguel A Arcones and Evarist Gine. "On the bootstrap of U and V statistics". In: *The Annals of Statistics* (1992), pp. 655–674 (cit. on p. 158).

[240] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000 (cit. on p. 159).

[241] Leon Cohen. *Time-frequency analysis*. Vol. 778. Prentice hall, 1995 (cit. on p. 165).

[242] Br Gold, AV Oppenheim, and CM Rader. "Theory and implementation of the discrete Hilbert transform". In: *Proc. Symposium on Computer Processing in Communications, Polytechnic Institute of Brooklyn*. 1969 (cit. on p. 165).

[243] Hendrikus Duifhuis. *Cochlear mechanics: introduction to a time domain analysis of the nonlinear cochlea*. Springer Science & Business Media, 2012 (cit. on pp. 167, 168).

[244] Andrew Bell. “A resonance approach to cochlear mechanics”. In: *PloS one* 7.11 (2012), e47918 (cit. on p. 167).

[245] Alberto Recio-Spinoso and William S Rhode. “Fast waves at the base of the cochlea”. In: *PloS one* 10.6 (2015), e0129556 (cit. on p. 167).